UNIVERSITÉ DE SHERBROOKE

Les jeunes face à l'éducation et au marché du travail : Enquête et étude de cas au Ghana

par
Anne Duplantier

Thèse présentée en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
Doctorat en économie du développement

Février 2021

UNIVERSITÉ DE SHERBROOKE

École de gestion

Les jeunes face à l'éducation et au marché du travail : Enquête et étude de cas au Ghana

Anne Duplantier

Cette thèse a été évaluée par un jury composé des personnes suivantes :

Claudia Champagne	Présidente de jury
Kim Lehrer	Co-directrice de recherche
Martino Pelli	Co-directeur de recherche
Patrick Richard	Évaluateur interne
Sonia Laszlo	Évaluatrice externe

Résumé

L'éducation est maintenant reconnue par les économistes comme ayant un rôle clé dans la croissance économique et le processus de développement d'un pays. D'ailleurs l'éducation primaire et secondaire ont été des objectifs prioritaires des acteurs du développement ces vingt dernières années, délaissant quelque peu les questions relatives à l'éducation supérieure dans les pays en développement. Cette thèse met en lumière les liens entre l'éducation et le marché du travail au Ghana. En particulier, le passage de l'éducation secondaire à l'éducation supérieure, ainsi que la transition vers le marché du travail sont étudiés à travers différents angles : la migration interne et les revenus espérés des jeunes Ghanéens. La thèse repose sur une base de données primaire collectée au Ghana et utilisée tout au long de ce travail.

Dans le premier chapitre, la base de données, l'expérience de la collecte ainsi que le système scolaire du Ghana sont présentés. De plus, ce chapitre apporte une description détaillée et approfondie de l'éducation secondaire et post-secondaire au Ghana ainsi que la situation des jeunes éduqués sur le marché du travail. Les méthodes descriptives utilisées ont recours à des graphiques et tableaux de statistiques et des tests de différence de coefficients sont menés afin d'étudier les différences par genre. Les résultats indiquent que les filles ont moins accès à l'école secondaire que les garçons, mais elles ont une plus grande probabilité de fréquenter une institution post-secondaire. Cependant, les filles ont tendance à se tourner vers des études post-secondaires moins longues et académiques que leurs congénères de sexe masculin. Concernant le marché du travail, les filles ont une plus forte probabilité d'être au chômage après leurs études. Il semble exister une prime à l'éducation au Ghana, puisque les individus ayant fréquenté une institution post-secondaire ont en moyenne des revenus plus élevés que les autres.

Le deuxième chapitre analyse les liens entre la migration interne des jeunes éduqués et le marché du travail. Son principal objectif est d'étudier le rôle des revenus espérés moyens régionaux et de leur variabilité dans la décision de changer de région de résidence des jeunes qui sortent de l'école secondaire. La méthodologie utilisée est un modèle de choix comportemental estimé par un logit mixte. Cette méthode permet de dépasser certaines limites du modèle logit standard telle que l'hypothèse d'indépendance des alternatives non pertinentes. Les résultats indiquent qu'un revenu moyen supérieur dans une autre région du pays augmente la probabilité de migrer dans cette région. Les différences régionales dans la variabilité des revenus sont aussi positivement liées à la migration interne. Ceci suggère que les jeunes éduqués migrants sont attirés par le risque ou bien qu'ils pensent pouvoir réussir mieux que les autres sur le marché du travail et être positionnés dans la fourchette haute des revenus. Enfin, les jeunes sortant de l'école secondaire qui sont nés

dans une région rurale ont une plus faible probabilité de migrer, alors que la probabilité de migrer est positivement liée à l'éducation de la mère et aux capacités individuelles du jeune.

Le troisième chapitre étudie le rôle des attentes dans les choix d'éducation post-secondaire au Ghana. L'objectif est d'analyser le lien entre les rendements espérés de l'éducation et les décisions de postuler et de fréquenter l'université. De plus, le rôle des capacités espérées et réelles est pris en compte. La méthodologie utilisée pour cela est un modèle linéaire de choix. Les résultats indiquent que les rendements espérés de l'éducation sont positivement liés à la décision de candidater à l'université. Cependant, une fois que l'on prend en compte les capacités réelles et estimées, les rendements espérés ne sont plus significativement liés aux décisions d'éducation post-secondaire. Les capacités réelles et perçues sont positivement liées à la probabilité de candidater à l'université ainsi que la décision de fréquenter l'université. Il est donc nécessaire et important de prendre en compte les capacités quand on étudie la décision de s'éduquer.

Finalement, le quatrième chapitre aborde la question technique du traitement de l'attrition qui touche la base de données en panel utilisée dans cette thèse. L'attrition et ses caractéristiques sont décrites et les déterminants de l'attrition sont analysés. Des tests de biais d'attrition sont mis en oeuvre et des corrections sont appliquées. Les résultats suggèrent que les individus sujets à l'attrition ont des caractéristiques différentes de ceux présents dans toutes les vagues. Ils sont en moyenne plus jeunes, ont une plus grande probabilité de venir de zones rurales et ont plus de frères et soeurs. De plus, ces personnes ont en général des parents moins éduqués et de moins bons résultats scolaires. Il est mis en évidence que les résultats d'intérêt du chapitre 3 souffrent d'un biais d'attrition. En effet, la probabilité de candidater et de fréquenter l'université est négativement liée à la probabilité d'être sujet à l'attrition. Ainsi une correction est appliquée aux résultats du Chapitre 3 et les résultats corrigés sont très proches de ceux non corrigés. Même si la base de données est sujette à l'attrition, les résultats restent pertinents et valides.

Mots clés : Économie du Développement ; Enquête ; Éducation ; Marché du Travail ; Migration Interne ; Ghana

Codes JEL : C8 ; I25 ; C25 ; R23 ; D84 ; I26 ; J24

Table des matières

Liste des tableaux

Liste des figures

Abréviations

ANOVA	Analyse of Variance
ASC	Alternative Specific Constant
ASS	Afrique Sub Saharienne
BECE	Basic Education Certificate Examination
BGLW	Becketti, Gould, Lillard and Welch
FGM	Fitzgerald, Gottschalk and Moffit
FE	Fixed Effect
GDP	Gross Domestic Product
GER	Gross Enrollment Ratio
GHS	Ghanaian Cedi
GLSS	Ghana Living Standard Survey
HIV	Human Immunodeficiency Viruses
IIA	Independance of Irrelevant Alternatives
IMR	Inverse of Mill's Ratio
IPW	Inverse Probability Weighting
JHS	Junior High School
LPM	Linear Probability Model
MMNL	Mixed Multinomial Logit
MNL	Multinomial Logit
OLS	Ordinary Least Squares
OMD	Objectif du Développement pour le Millénaire
PPP	Purchasing Power Parity
PSE	Post-Secondary Education
PSID	Panel Study of Income Dynamics
PSM	Propensity Score Matching
RDD	Regression Discontinuity Design
SHS	Senior High School
SSA	Sub Saharan Africa
TBS	Taux brut de scolarisation
TNS	Taux net de scolarisation
UNDP	United Nations Development Programme
USA	United States of America

USD	United States Dollar
WASSCE	West African Senior School Certificate Examination
WDI	World Development Indicators
WLS	Weighted Least Squares

Remerciements

Ces années de doctorat ont été particulièrement intenses et riches, à la fois sur le plan professionnel que personnel. La découverte et l'apprivoisement du monde de la recherche tout autant que celui du Québec m'ont donnée l'occasion d'apprendre beaucoup intellectuellement, culturellement, personnellement, et de découvrir de nouveaux possibles. Tout au long de ce cheminement qui fut parfois difficile, j'ai eu la chance de pouvoir compter sur de nombreuses personnes que j'aimerai ici remercier.

Je tiens à exprimer une sincère reconnaissance envers ma directrice de thèse, Kim Lehrer, pour son encadrement rigoureux et exigent ainsi que son accompagnement durant toutes ces années. Elle m'a fait confiance dès les premiers instants en m'embarquant dans son projet d'enquête au Ghana, ce qui m'a permis de réaliser mon rêve d'étudiante en économie du développement. Elle m'a transmis son goût pour le terrain et m'a permis d'acquérir une solide expérience de traitement de données d'enquête en tant qu'assistante de recherche. Lors des moments difficiles, elle a fait preuve d'une grande patience, n'a jamais semblé douter de mes capacités et a su adapter ses exigences quand c'était nécessaire. Ses encouragements et sa compréhension ont grandement contribué à l'aboutissement de ce projet de doctorat. Un grand merci également à Martino Pelli pour sa confiance et son accompagnement lors des dernières étapes vers la soutenance.

Je souhaite remercier Valérie Vierstraete et Mario Fortin qui m'ont fait confiance et m'ont donné l'occasion d'enseigner à plusieurs reprises. Cette expérience d'enseignement a développé mon goût pour l'enseignement et a constitué une source d'épanouissement et de satisfaction. Un merci tout particulier pour Valérie Vierstraete qui m'a toujours ouvert sa porte lorsque j'avais besoin d'une oreille bienveillante et réconfortante. J'ai aussi une pensée pour certains professeurs et chercheurs qui m'ont aidé à m'engager sur ce chemin d'économiste. Parmi eux, Simone Bertoli et Romain Beccucci m'ont particulièrement orientée et ont cru en moi à certains tournants de mon parcours.

Finalement, dans mon quotidien d'étudiante exilée, un certain nombre de phares m'ont permis de tenir le cap et d'arriver à bon port : Le soutien et l'accompagnement de mes amis de Sherbrooke qui resteront un des magnifiques cadeaux de cette aventure. L'affection et la présence de mes amis de France qui sans faillir ont traversé l'océan et les années. La confiance, les encouragements et l'amour de mes parents et mes frères et soeur. L'épanouissement et le ressourcement provenant du chant, que j'ai toujours pu continuer à pratiquer avec passion. J'ai également une pensée émue pour ma grand mère, qui a renoncé à son projet de thèse afin que mon grand père puisse mener à bout la sienne. Un immense merci, enfin, à Anicet avec qui je partage cette vie d'économiste et bien plus encore, à travers les continents.

INTRODUCTION GÉNÉRALE

Le rôle clé de l'éducation dans la croissance économique et le processus de développement d'un pays a d'abord été mis en lumière par les travaux fondateurs de Schultz (1961) et Becker (1964). En effet, ces auteurs sont les premiers économistes à considérer les connaissances et compétences comme du capital dans lequel les individus investissent. Ils soutiennent que l'investissement en capital humain est une des principales sources de la croissance économique. Une vingtaine d'années après, Hicks (1980) se penche lui aussi sur le lien entre croissance économique et développement humain dans 83 pays en développement entre 1960 et 1977. Il considère l'alphabétisation comme une mesure de développement de l'éducation et l'espérance de vie comme une autre mesure de développement humain. Hicks (1980) trouve que les 12 pays en développement dont la croissance est la plus forte sont ceux dont les niveaux d'alphabétisation et d'espérance de vie sont les plus élevés. De plus, trois facteurs expliquant 60% de la croissance dans les pays en développement sur cette période sont mis en évidence par : le taux d'investissement, le taux de croissance des importations et le niveau de développement des ressources humaines en 1960.

Ces vingt dernières années, l'éducation primaire et secondaire ont été des objectifs prioritaires des acteurs du développement. En effet, le deuxième Objectif du Développement pour le Millénaire (OMD) était d'assurer l'éducation primaire pour tous en 2015. Selon les Nations Unies (2015), le taux net de scolarisation [1] (TNS) au primaire est ainsi passé de 83% en 2000 à 91% en 2015 dans les pays en développement. Dans toutes les régions du monde, des améliorations ont été réalisées mais c'est en Afrique Sub-Saharienne (ASS) que les avancées en matière de scolarisation primaire ont été les plus fortes (Nations Unies, 2015). En effet, le taux net de scolarisation au primaire a augmenté de 20 points de pourcentage entre 2000 et 2015 (contre une augmentation de 8 points de pourcentage entre 1990 et 2000), et est ainsi passé de 60% en 2000 à 80% en 2015.

S'agissant de l'éducation secondaire les progrès sont un peu plus lents. En 2000, seulement 55% des jeunes en âge d'aller à l'école secondaire dans le monde y étaient effectivement inscrits. Ce chiffre a atteint 66% en 2018, ce qui dénote une amélioration mais aussi les efforts qu'il reste à faire pour assurer l'accès à l'éducation secondaire pour tous. D'ailleurs, rendre universelle l'éducation de base jusqu'au niveau secondaire inférieur est un des Objectif de Développement Durable pour 2030. Dans les pays en développement, les avancées en matière d'éducation secondaire varient selon les régions. En Amérique Latine, Asie de l'Est et Moyen Orient-Afrique du Nord, le taux net de scolarisation au niveau secondaire est passé respectivement de 66%, 64%, 61% en 2000 à 78%,

1. Le taux net de scolarisation est la proportion de personnes scolarisées correspondant à la tranche d'âge officielle de la scolarisation en question, sur toute la population de personnes en âge d'être scolarisées à ce niveau d'éducation.

79%, 73% en 2018 (World Bank, 2020). Les deux régions les plus en retard d'après ce rapport de la Banque Mondiale sont l'Asie du Sud (39% en 2000 et 61% en 2018) et l'ASS dont le taux est passé de 21% en 2000 à 36% en 2018. Les chiffres sont un peu plus élevés si l'on regarde au niveau secondaire inférieur : le taux net de scolarisation était de 41% en 2000 contre 66% en 2014 en ASS (Bashir et al., 2018).

En matière d'éducation, l'objectif prioritaire des institutions internationales dans les pays en développement est l'éducation de base. En conséquent, l'accent a moins été mis sur l'accès et la qualité de l'éducation supérieure. Pourtant, investir dans l'éducation supérieure est important pour le processus de développement. La Banque Mondiale définie l'enseignement supérieur comme toute éducation post-secondaire, incluant les universités publiques et privées, les instituts de formation technique et les écoles professionnelles. Les personnes les plus éduquées participent particulièrement à la croissance et la réduction de la pauvreté (Bloom et al., 2006). En effet, elles ont de meilleures compétences, gagnent des salaires plus élevés et sont plus aptes à gérer les chocs. La société bénéficie également de l'enseignement supérieur puisque les citoyens les plus instruits prennent davantage soin de leur santé, sont plus sensibles à la cause environnementale ainsi qu'à la participation civique.

De plus, des revenus plus élevés augmentent les potentielles recettes fiscales du gouvernement. Montenegro et Patrinos (2014) soulignent qu'en général, les rendements économiques privés de l'éducation post-secondaire sont supérieurs à ceux de l'éducation primaire et secondaire : 14.6% contre 11.5% et 6.8% respectivement. En ASS, les rendements des investissements dans l'enseignement supérieur sont les plus élevés du monde avec 21%, sachant que la moyenne mondiale est de 14.6% (Montenegro et Patrinos, 2014). Au Ghana, les rendements de l'éducation supérieure étaient estimés à 28.7% en 2012, avec une légère différence entre hommes (26.6%) et femmes (24.8%) d'après Montenegro et Patrinos (2014).

Darvas et al. (2017) soulignent la croissance de la fréquentation de l'enseignement supérieur en ASS. Le taux de croissance annuel du taux brut de scolarisation [2] (TBS) dans l'enseignement supérieur en ASS était de 4.3% entre 1970 et 2013 contre 2.8% pour la moyenne mondiale (Darvas et al., 2017). Cependant dans les pays en développement, fréquenter une institution post-secondaire n'est pas accessible à tous et les étudiants se heurtent à de nombreux obstacles. Le TBS d'ASS reste le plus faible comparé aux autres régions du monde, puisqu'il était de 10% en 2013 contre 30% en moyenne dans le monde (Darvas et al., 2017). Selon Psacharopoulos et Patrinos (2018), la durée moyenne de scolarisation est de 5,2 années en ASS, contre 8 ans dans le monde et 9,5 années dans les économies avancées.

2. Le taux brut de scolarisation est défini comme le nombre de personnes scolarisées à un certain niveau d'éducation sur la population totale du groupe d'âge officiel associé à ce niveau d'éducation.

D'après la Banque Mondiale, seuls 7,9% des habitants d'ASS étaient inscrits dans des établissements d'enseignement supérieur en 2017, contre 37,9% en moyenne dans le monde (World Bank, 2019). De plus, le rapport de la Banque Mondiale (2017) "Sharing higher education in SSA" souligne qu'il manque en général en ASS un méchanisme efficace permettant d'évaluer les compétences des diplômés et de combler l'inadéquation des compétences sur le marché du travail. Dans certains pays comme le Cameroun, la Côte d'Ivoire, Madagascar, la Mauritanie, le Niger, le Nigeria, le Sénégal, la Tanzanie et l'Uganda, les diplômés de l'enseignement post-secondaire ont un taux de chômage supérieur aux jeunes qui ont seulement un niveau d'éducation primaire ou secondaire (Darvas et al., 2017).

Un des défis de l'éducation en ASS est de développer un système d'enseignement secondaire et post-secondaire qui ne se contente pas de fournir des connaissances académiques mais également des compétences professionnelles ainsi qu'un esprit civique, comme le soulignent Gonzalez et Rawls (2007) dans le rapport de la Conférence Afrique 2007 de McGill. Le faible taux d'inscription dans l'enseignement supérieur en Afrique peut s'expliquer par plusieurs facteurs. Une première hypothèse met en avant les contraintes de crédit qui empêchent le financement des études post-secondaires, même si certains étudiants ont parfois accès à des bourses. Une deuxième explication est l'accès limité à l'éducation post-secondaire qui viendrait de l'incapacité des certains étudiants à atteindre le niveau suffisant et d'un trop faible nombre de places offertes dans des établissements d'enseignement supérieur par rapport à la demande.

Un troisième facteur est le fait que les étudiants pensent que les rendements de l'éducation supérieure sont trop faibles et que la probabilité d'obtenir un emploi est risquée. Il est possible que les étudiants des pays en développement fondent leurs décisions sur des informations erronées ou imparfaites. Une quatrième raison vient du fait que les incitations à l'éducation sont trop faibles et les coûts d'opportunité trop élevés : il n'y a pas de gain direct à poursuivre la scolarité après le primaire. Enfin, le dernier facteur de choix en matière d'éducation sont les préférences. Les individus essaient autant que possible de choisir leur profession et leur scolarité en fonction de ce qu'ils aiment faire et de ce qu'ils savent faire.

Le rapport de la Banque Mondiale "Sharing higher education's promise beyond the few in Sub-Saharan Africa" rapporte que le système d'enseignement supérieur du Ghana est l'une des réussites d'ASS (Darvas et al., 2017). Le taux d'inscription dans l'éducation tertiaire au Ghana est relativement élevé comparé à la sous-région : 1370 inscriptions pour 100 000 habitants (Darvas et al., 2017). De plus, le taux d'inscription brut dans l'enseignement supérieur était de 16,2% en 2017 (et 12,1% en 2011) d'après les indicateurs de développement mondial de la Banque mondiale. Toutefois, il y a de fortes disparités entre les jeunes des ménages aisés (TBS de 13% pour le quintile le plus riche) et ceux des ménages pauvres (TBS de 1,4% pour les 40% les plus pauvres). Les coûts

de l'enseignement supérieur au Ghana sont considérés comme relativement "raisonnables" d'après Darvas et al. (2017). Le ratio coût par étudiant dans l'enseignement supérieur-coût par étudiant dans l'enseignement primaire est de 3,9 au Ghana, contre 224 au Malawi ou 13 en Guinée et en Ouganda.

Psacharopoulos et Schultz (1984) analysent la contribution de l'éducation à la croissance économique dans 29 pays en développement au milieu des années 1980. Ils trouvent que l'éducation contribue à 23% de la croissance annuelle Ghanéenne à cette période. De nos jours, le lien entre éducation et développement est toujours étudié par les économistes. Selon Montenegro et Patrinos (2014), les rendements d'une année d'éducation sont les plus élevés en ASS : 12.4% pour la sous-région et 12.2% au Ghana en 2012, considérablement supérieur à la moyenne mondiale qui est à 9.7%. Cependant, les liens entre éducation et marché du travail ne sont pas toujours si simples. D'après Garcia et Jean (2008), il y aurait une relation en cloche entre le niveau d'éducation et le chômage au Ghana. En effet, en 2005, le taux de chômage était de 70% pour les jeunes sans éducation, de 75% pour les jeunes avec le niveau primaire, 80% pour un niveau d'éducation secondaire alors que le taux de chômage des jeunes avec éducation supérieure était de 40%.

Ainsi, l'objectif général de cette thèse est de mieux comprendre les liens entre éducation et marché du travail au Ghana. Pour cela, une base de données primaire collectée auprès de jeunes éduqués est utilisée dans les quatre chapitres de cette thèse. Cette base de données en panel suit les mêmes étudiants à travers le temps entre 2011 et 2018. Afin de comprendre la représentativité de l'échantillon de la base de données utilisée dans cette thèse, j'utilise la sixième vague de l'enquête Living Standard Measurement Survey (Ghana Living Standard Survey 6, noté GLSS6) qui est représentative de la population du Ghana. Les jeunes entre 18 et 28 ans ayant un niveau d'éducation secondaire (la population étudiée dans cette thèse) représentent environ 5% de l'échantillon GLSS6. De plus, 23,7% des jeunes de 18 à 28 ans de GLSS6 ont atteint le niveau secondaire comme plus haut niveau d'éducation.

Le premier chapitre est un chapitre descriptif et de mise en contexte qui recouvre plusieurs objectifs. Tout d'abord, ce chapitre permet de présenter la base de données en panel que j'ai contribué à collecter ainsi que les choix méthodologiques nécessaires pour la collecte, l'expérience de terrain et la gestion de la base de données. Le contexte du Ghana, pays sur lequel portent les quatre chapitres de la thèse, et son système scolaire sont également abordés. De plus, le premier chapitre apporte une description détaillée et approfondie de l'éducation secondaire et post-secondaire au Ghana ainsi que la situation des jeunes éduqués sur le marché du travail. Les méthodes descriptives utilisées ont recours à des graphiques et tableaux de statistiques. De plus, des tests de différence de coefficients sont menés afin d'étudier les différences par genre.

La principale contribution de ce chapitre est d'apporter de l'information sur une population peu

étudiée d'un pays d'ASS en décrivant la situation des jeunes Ghanéens dans leur scolarité secondaire et post-secondaire. En particulier, la base de données apporte des informations détaillées sur les résultats scolaires au secondaire et l'accès des étudiants aux études post-secondaires. De plus, leur transition et intégration sur le marché du travail est analysée. Enfin, une attention particulière est apportée sur les différences selon le genre afin de comprendre si les résultats d'éducation et de marché du travail varient selon que l'on est une fille ou un garçon.

Cette analyse descriptive permet de mettre en évidence certaines différences selon le genre. Les filles ont moins accès à l'école secondaire que les garçons, mais elles ont une plus grande probabilité de fréquenter une institution post-secondaire que les garçons. Cependant, il y a des différences selon le genre dans le type d'institution fréquentée. Les filles ont une plus forte propension à faire des études d'infirmière alors que les garçons ont plus de chance d'aller à l'université que les filles. De plus, les filles à l'école secondaire viennent en moyenne de ménages moins pauvres et avec un meilleur niveau de vie que les garçons. Concernant les résultats touchant au marché du travail, les filles ont une plus forte probabilité d'être au chômage après leurs études. On trouve aussi que les individus qui ont fréquenté une institution post-secondaire ont en moyenne des revenus plus élevés que les autres.

Le deuxième chapitre analyse les liens entre la migration interne des jeunes Ghanéens éduqués et le marché du travail. Son principal objectif est d'étudier le rôle des revenus espérés moyens régionaux et de leur variabilité dans la décision de changer de région de résidence des jeunes qui sortent de l'école secondaire. Pour cela, une partie de la base de données en panel présentée dans le Chapitre 1 est utilisée. Ce sous-échantillon est constitué uniquement d'individus qui ne sont ni à l'école secondaire ni étudiant post-secondaire au moment de l'enquête afin de capter les déplacements des jeunes qui sont potentiellement sur le marché du travail. La méthodologie utilisée est un modèle de choix comportemental estimé par un logit mixte. Cette méthode permet de dépasser certaines limites du modèle logit standard telle que l'hypothèse d'indépendance des alternatives non pertinentes.

Ce chapitre contribue à la littérature sur la migration interne dans les pays d'Afrique de l'Ouest en considérant la migration comme une décision individuelle et en choisissant comme population d'étude les jeunes éduqués. En effet, cette population est d'une importance cruciale pour le développement d'un pays mais souvent peu étudiée par manque de données apropriées. De plus, au lieu de se concentrer uniquement sur les différences de revenus moyens comme déterminant de la migration, le risque est pris en compte en intégrant les différences de variations de revenus entre les régions.

Les résultats indiquent qu'un revenu moyen supérieur dans une autre région augmente la probabilité que les jeunes éduqués se déplacent dans cette région. Les différences régionales dans la variabilité

des revenus sont aussi positivement liées à la migration interne. Ceci suggère que les jeunes éduqués migrants sont attirés par le risque ou bien qu'ils pensent pouvoir réussir mieux que les autres sur le marché du travail et être positionnés dans la fourchette haute des revenus. Enfin, les jeunes sortant de l'école secondaire qui sont nés dans une région rurale ont une plus faible probabilité de migrer, alors que la probabilité de migrer est positivement liée à l'éducation de la mère et aux capacités individuelles du jeune.

Le Chapitre 3 porte sur le rôle des attentes dans les choix d'éducation post-secondaire au Ghana. L'objectif est d'étudier le lien entre les rendements espérés de l'éducation et les décisions de postuler et de fréquenter l'université. De plus, le rôle des capacités espérées et réelles est analysé. Un sous-échantillon de la base de données en panel présentée dans le Chapitre 1 est utilisé pour conduire cette analyse. En particulier, la dimension panel de la base est utilisée afin d'obtenir pour chaque individu les rendements espérés de l'université lorsqu'il était encore à l'école secondaire, mais aussi le choix d'institution post-secondaire fait après être sorti de l'école secondaire. La méthodologie utilisée ici est un modèle linéaire de choix.

Ce travail contribue à la littérature en décomposant le processus de choix d'éducation post-secondaire en deux étapes : d'une part la candidature et de l'autre la fréquentation. Habituellement, seule la fréquentation est étudiée. De plus, au lieu de se concentrer uniquement sur les rendements espérés de l'éducation comme déterminant des choix d'éducation, deux nouvelles composantes sont ajoutées : les capacités réelles de l'individu, mesurées par les résultats académiques secondaires, ainsi que les capacités perçues, mesurées avant que l'individu connaisse ses résultats académiques. Finalement, cette base de données permettant d'analyser les attentes en matière d'éducation et de marché du travail des jeunes éduqués permet de combler le manque de données et de connaissances à ce sujet dans la région d'ASS.

Les résultats indiquent que les rendements espérés de l'éducation sont positivement liés à la décision de postuler à l'université. Cependant, une fois que l'on prend en compte les capacités réelles et estimées, les rendements espérés ne sont plus significativement liés aux décisions d'éducation post-secondaire. Les capacités réelles et perçues sont positivement liées à la probabilité de postuler à l'université ainsi que la décision de fréquenter l'université. Il est donc nécessaire et important de prendre en compte les capacités quand on étudie la décision de s'éduquer.

Le Chapitre 4 aborde la question technique du traitement de l'attrition. L'objectif est de conduire une analyse d'attrition complète à partir de la base de données de panel utilisée dans cette thèse. Afin de comprendre l'ampleur et les caractéristiques de l'attrition, une première partie de ce chapitre porte sur la description de l'attrition ainsi que les caractéristiques des individus sujets à l'attrition. Dans un deuxième temps, les déterminants de l'attrition sont analysés. Enfin des tests sont mis en oeuvre afin de comprendre si les résultats d'intérêt des chapitres précédents souffrent d'un

biais d'attrition, et le cas échéant, une correction est appliquée.

Différentes méthodes sont employées pour chaque étape. Afin de comprendre quelles caractéristiques diffèrent chez les individus sujets à l'attrition par rapport aux autres individus de l'échantillon, deux méthodes sont utilisées. Tout d'abord des tests (t-test et Chi-deux test) permettent de tester les différences de moyenne entre les deux groupes. Ensuite, la probabilité d'être touché par l'attrition est estimée par les moindres carrés ordinaires afin de contrôler pour les effets de cohortes. Ensuite, deux types de tests couramment utilisés dans la littérature sur l'attrition sont mis en oeuvre afin de vérifier si les résultats d'intérêts souffrent de biais d'attrition. Ces tests sont ceux développés par Fitzgerald et al. (1998) d'une part, et Becketti et al. (1988) d'autre part. Enfin, la correction est appliquée en utilisant deux méthodes : celles des moindres carrés pondérés et celle de la procédure d'Heckman (Heckman, 1979).

Bien que toute analyse d'attrition soit spécifique à la base de données et aux résultats d'intérêts étudiés, ce chapitre apporte quelques contributions à la littérature. Tout d'abord, la base de données qui comporte plusieurs vagues d'enquête permet de distinguer quatre types d'attrition au lieu de deux habituellement. De plus, alors que la grande majorité des travaux sur l'attrition se concentrent sur l'attrition venant des caractéristiques observables, les spécificités de la base de données nous permettent d'étudier également l'attrition venant des caractéristiques inobservables.

Les résultats suggèrent que les individus sujets à l'attrition ont des caractéristiques différentes de ceux présents à toutes les vagues. Ces individus sont en moyenne plus jeunes, ont une plus grande probabilité de venir de zones rurales et ont en moyenne plus de frères et soeurs. De plus, les individus sujets à l'attrition ont en général des parents moins éduqués et de moins bons résultats scolaires. Les résultats des tests montrent que les résultats d'intérêt du Chapitre 3 souffrent d'un biais d'attrition. En effet, la probabilité de postuler et de fréquenter l'université est négativement liée à la probabilité d'être sujet à l'attrition. Ainsi une correction est appliquée aux résultats du Chapitre 3 et les résultats corrigés sont très proches de ceux non corrigés. Même si la base de données est sujette à l'attrition, les résultats restent pertinents et valides.

Ainsi, cette thèse contribue à la connaissance en économie du développement en apportant des éclairages sur la relation entre l'éducation et le marché du travail dans un pays en développement. Pour cela, j'explore la situation des jeunes qui ont un niveau d'éducation secondaire sur le marché du travail, ainsi que lors de leur parcours post-secondaire (Chapitre 1). La relation entre la migration interne des jeunes éduqués et les revenus moyens des différentes régions du pays est abordée dans le Chapitre 2. De plus, la question des attentes des jeunes concernant les rendements de l'éducation est approfondie afin de comprendre si ces attentes ont un lien avec les choix d'éducation post-secondaire (Chapitre 3). Finalement, la robustesse de ces analyses à l'attrition est éprouvée dans le Chapitre 4.

CHAPITRE 1
YOUTH EDUCATION AND LABOR MARKET OUTCOMES IN GHANA : EVIDENCE AND DESCRIPTIVE STATISTICS FROM PRIMARY DATA

Résumé

Dans les pays en développement, peu d'information est disponible concernant l'éducation post-secondaire et la transition des jeunes sur le marché du travail. Pourtant, les jeunes éduqués sont un élément d'avenir et crucial pour un pays en développement. Ce chapitre permet d'aborder de façon détaillée ces questions en décrivant la situation des jeunes face à l'éducation secondaire et post-secondaire, ainsi que leur insertion sur le marché du travail. De plus, la base de données, les choix méthodologiques concernant la collecte ainsi que le système scolaire du Ghana sont présentés. Les résultats indiquent que les filles ont moins accès à l'école secondaire que les garçons, mais elles ont une plus grande probabilité de fréquenter une institution post-secondaire. Cependant, les filles ont tendance à se tourner vers des études post-secondaires moins longues et académiques que leurs congénères de sexe masculin. Concernant le marché du travail, les filles ont une plus forte probabilité d'être au chômage après leurs études. Il semble exister une prime à l'éducation au Ghana, puisque les individus ayant fréquenté une institution post-secondaire ont en moyenne des revenus plus élevés que les autres.

1.1 INTRODUCTION

Although the young and growing working-age population in Sub-Saharan Africa (SSA) represents a substantial economic advantage, unemployment and the transition from education to the labor market is a crucial challenge, limiting the ability of SSA countries to take full advantage of their young population. According to Ackah Baidoo (2016), youths represent three-fifths of the unemployed people in Africa. In SSA in particular, 18% of youths are unemployed (ibid.). The weak quality of education, measured by cognitive skills development, appears to be one of the issues (Fox et al., 2016). The mismatch between education and the labor market also plays a role. For Pitan and Adedeji (2012), university graduates in Nigeria are not adequately prepared for work with respect to the demand of skills in the labor market. As highlighted in Filmer and Fox (2014), an effective allocation of the skills in a country allows for productive employment, reduces unemployment rates and contributes to the development process. According to the authors, schooling is associated with greater productivity and earnings, and strongly influences the sector in which people work. A higher level of education facilitates entry into non-farm household enterprises and modern wage jobs.

Similar to other SSA countries, Ghana faces the same challenge with significant unemployment, in particular for youth. In 2011, the youth unemployment rate in Ghana was 65% according to the World Bank. [1] According to Avura and Ulzen-Appiah (2016), Ghanaians aged 15-24 are much less likely to be working than adults aged 25-65, due mainly to : (1) a deficit in job-relevant skills, (2) a lack of job search experience, (3) difficulties in obtaining information about career options and (4) a lack of access to finance for youth. Moreover, Ghana Statistical Service (2012) highlights that the majority (58.9%) of the unemployed persons are first time job seekers.

However, there is a lack of evidence, data and information to study and understand youth education, unemployment, and their transition to the labor market. To the best of our knowledge, the Young Lives project at the University of Oxford is one of the few datasets covering issues of poverty, education, health, employment and opportunities of youth in three developing regions (South Asia, East Africa and Latin America). In order to understand youth poverty in developing countries, the project follows 12 000 children in Ethiopia, India, Peru and Vietnam over 18 years. The sample includes two cohorts of children, an older cohort born in 1994-95 and a younger cohort born in 2001-02, allowing comparison of youth at the same age at different moments. The survey tackles various subjects as poverty, health, education, gender, and transition toward labor market. In particular, the longitudinal data enables better understanding of which skills are crucial for entering into the labor market, whether the education choices of youth match with their aspirations and what are

1. https ://web.archive.org/web/20120213132827/http ://www.ghanatoghana.com/Ghanahomepage/world-bank-assist-national-youth-employment-programme

their expectations about higher education.

Assaad et al. (2016) use labor market panel surveys to understand the issue of labor market insertion and youth unemployment in three countries of the Middle East and North Africa. They find that education level of a father impacts the probability of their children to be unemployed. Moreover, the longer a man is unemployed, the greater the probability of his leaving the country. The results also suggest that educated young women aspire to work but lack the requisite mobility to move to where the jobs are. Tiongson and Fares (2007) study the labor market transition and long-term effects of youth unemployment by using panel data in Bosnia and Herzegovina. The authors find that youth unemployment is twice the national average and youth with higher education attainment are less likely to be jobless. In Colombia, Attanasio et al. (2007) examine the effects of a randomized training on the labor market outcomes. The results suggest that trained women have a higher probability of being employed, earn higher wages and work more days.

This dissertation fills a gap in these studies with data by focusing on education and labor market outcomes of youth in Ghana. The project, called Ghana Opportunities for Transitioning Senior High School Students, builds a panel database thanks to survey following the same students over time (between 2011 and 2018). It brings crucial and detailed information on secondary and post-secondary schooling, as well as the transition towards the labor market, allowing the analysis of youth education and labor market outcomes in a Sub-Saharan country. Thanks to this dataset, researchers will be able to understand the long-term patterns toward education and labor market. Bringing detailed statistics about secondary education (the success or not at the exam of end of secondary) and the process to access post-secondary education (the application process) will shed light on the strengths and failures of the Ghanaian education system in order to help to improve it. The survey also brings information about the search for jobs and the first years of active life after school. By knowing the number of jobs youth have at the same time, the worked hours and the estimated income, researchers can sketch the situation of the young Ghanaian who attended Senior High School (SHS). For all these reasons, the survey contributes to the knowledge in development economics.

In this chapter, I present the panel database which I contributed to collect, the methodology used during the data collection and an overview of the main characteristics of the sample. The chapter is organized as follows. Section 2 presents a rapid overview of Ghana, the Ghanaian school system and the panel survey. Then, methodological choices made for the survey are discussed in Section 3. Section 4 addresses the field experience and the data management. A description of the main characteristics of the sample in terms of personal characteristics, as well as education and labor market outcomes, is presented in Section 5. The last section concludes this descriptive analysis.

1.2 CONTEXT

1.2.1 Overview of the Ghana

Ghana is a SSA country of 238 540 square kilometres, bordered by Côte d'Ivoire, Togo and Burkina Faso. Several ethnic groups co-exist in the population, with the Akans being the largest group (48%), followed by the Mole-Dagbami (17%) and the Ewe (14%) (Ghana Statistical Service et al., 2015). Christians represent 70% of the population and Muslims 18%, while only the Northern region is predominantly Muslim at 60% (ibid.). English is the official language but there are a number of other language groups (at least 69 dialects).

Ghana's population was estimated at 29.7 million in 2018 according to World Bank (2019). In 2015, Ashanti, Eastern and Greater Accra regions accounted for 50% of the population, while the Upper East region was the least populated, with only 2% of the country population (Ghana Statistical Service et al., 2015). In 2015, 51% of the population was living in urban areas (ibid.). However, there were wide regional disparities. The urbanization rate was 90% in Greater Accra, 60% in Ashanti but only 16% in Upper West (ibid.).

In 2015, 40% of the population was under 14 years old (Ghana Statistical Service et al., 2015). In the 2020 Human Development Report from the United Nations Development Programme (UNDP), the Ghanian human development index (HDI) was estimated at 0.61 (in 2019 value) on a scale of 0-1 (United Nations Development Programme, 2020). The different components of this index are : the life expectancy at birth (64.1 years in Ghana), the expected years of schooling (11.5 in Ghana), the mean years of schooling (7.3 in Ghana) and the gross national income per capita (5,269 purchasing power parity (PPP) $). It situates the Ghana in the "Medium Human Development" group at the rank 138 over 189 countries. According to the 2010 census, 19% of the population is considered as an internal migrant (born in a region other than the region of residence) and about 1% of the population lives abroad (Ghana Statistical Service, 2012). In addition, 40% of Ghanaian households report having at least one member who is a migrant (Ackah and Medvedev, 2012).

Between 2008 and 2012, the average growth rate was 8.7% per year. Since then, the economy has slowed considerably, reaching an annual growth rate of 3.6% in 2017 (Ministry of Education Ghana, 2018b). Ghana has recently made the transition to lower-middle-income country status. The poverty rate, i.e. the percentage of the population living below the national poverty line, was 23.4% in 2016 (World Bank, 2019). In terms of inequality, the Gini coefficient is 0.41 and has remained relatively stable since 2006. Ministry of Education Ghana (2018b) points out that poverty is highest in the north-west of Ghana (the least populated regions) while the south-east of the country, the most populated part, is the least affected by poverty.

In 2019, services sector represented almost half of gross domestic product (GDP) (47.2%), industry

for a little more than a third (34.2%) and agriculture for 18.5% (Ghana Statistical Service, 2019). A large part of the working population is self-employed (65%), while employees and family workers account for 18% and 11.5% of the working population, respectively. However, the proportion of employees in urban areas (35.1%) is about three times higher than their counterparts in rural areas (10.2%). In addition, 45.8% of all households in Ghana are agricultural households, i.e. at least one member is engaged in one agricultural activity such as crop farming, tree planting, fish farming or animal rearing (Ghana Statistical Service, 2012).

The vast majority of employed people work in the informal private sector (86%) while only 7% work in the formal private sector and 6% in the public sector (Ghana Statistical Service, 2012). There are some disparities in the probability to work in the informal sector. Indeed, 61.5% of the urban workers are engaged in the informal sector compared to 23.3% of their rural counterparts (Ghana Statistical Service, 2012). Moreover, women are more likely to be engaged in the informal sector than men (47.8% versus 35.5%). The informal workers are largely self-employed persons such as farmers, traders, food processors and artisans. According to Osei-Boateng and Ampratwum (2011), the inability of the formal private sector to generate jobs in their required quantities has pushed many into the informal sector. Another reason of the prevalence of the informal sector is the absence of unemployment benefits that makes informal activities a survival strategy or an alternative to generate income for many Ghanaians (Osei-Boateng and Ampratwum, 2011; Poku-Boansi and Afrane, 2016). A large number of informal sector workers in Ghana are trapped in poverty as they do not earn enough to lift themselves and their families out of poverty. Moreover, they often lack crucial skills and technology which affect their productivity (Osei-Boateng and Ampratwum, 2011).

Beside informal jobs, the Ghanaian labor market faces issues of unemployment. The concept of unemployment is not easy to capture in a development economy as it refers to those who work in the formal economy. In consequence, only a small proportion of the labor force is officially recorded as unemployed (Poku-Boansi and Afrane, 2016). Since many of the unemployed persons search for a job in the informal sector, they are not register with the Labor Department. Thus, underemployment -working less than 35 hours per week in the main job- is a more visible phenomenon than unemployment and is generally characterized by low-productivity and low-income rates. In 2012/2013 33% of the active population was considered as underemployed people (Ghana Statistical Service, 2014a). Regarding unemployment, it concerns 5% of the active population and the majority (58.9%) of them are first time job seekers (Ghana Statistical Service, 2012).

However, there are disparities inside the population. Indeed, women have a greater probability to be unemployed than men and the 15-25 years age group has the highest unemployment rate (Ghana Statistical Service, 2014a). In addition, there are spatial disparities as the urban population is more

likely to be unemployed than the rural dwellers. According to Baah-Boateng (2013), urban life attracts more than agriculture sector in the rural areas. This pushes many people into the cities in search for better-paid jobs that are however often not readily available. Indeed, Greater Accra and Ashanti are the region registering the highest proportion of urban unemployed population (Poku-Boansi and Afrane, 2016). An explication is that Greater Accra and Ashanti are destinations for migrants from other regions because of the concentration of economic, education, health, social and other infrastructures. Unemployment rates in Upper East and Upper West regions are also higher than the national average probably because of limited opportunities for their inhabitants.

Baah-Boateng (2013) analyzes the determinants of unemployment in Ghana. On the demand-side factors, there is a divergence between growth of real GDP (5.2% per year between 1984 and 2010) and growth of employment (2.5% per year over the same period). This conducts to a decline in average employment elasticity of national output, indicating low employment content of economic growth. According to Baah-Boateng (2013), this phenomenon is largely explained by the slow growth of the high labor absorption sectors (agriculture and manufacturing) on one hand and higher growth of low employment generating sectors (such as mining and finance). Moreover, the empirical results from Baah-Boateng (2013) show that selective job seekers are more likely to be unemployed than others. Indeed, individuals who search for a specific type of employment and for a full-time job have a higher rate of unemployment. This reflects the limited job opportunities compared to the job desire of job seekers. A higher reservation wage is also found to increase the probability to be unemployed (Baah-Boateng, 2013). This is an indication of the high expectation of job seekers relative to the reality in the labor market. Finally, there is also an issue of information on available jobs and financial constraints, in particular among the poorest who have limited search abilities.

On the supply-side of the labor market, Baah-Boateng (2013) highlights an increasing labor force relative to lower employment opportunities. In addition, the low quality of the labor force or its mismatch with the skill requirements are pointed out by Baah-Boateng (2013). Indeed, in 2010, less than 25% of the working age population has acquired at least secondary education while more than 25% has no formal education. Additionally, about half of the working age population have acquired just basic education which only enables them to read and write with no employable skills to secure employment in the formal segment of the labor market. However, there is an improvement in the quality of the labor force over the years. However, the relationship between education and employment is not clear. Sackey and Osei (2006) found that basic and secondary levels of education are associated with relatively more unemployment in Ghana. Indeed, there is a mismatch between the number of educated job seekers and the availability of jobs that require such skills. This is due to the slow growth of the formal sector employment on one hand, and the unwillingness of the educated people to accept the jobs available in the agriculture or informal sectors.

Empirical results from Baah-Boateng (2013) confirm that youth unemployment is a particular challenge. Ghanaian youths face discriminatory practices of employers, but they also have lower labor market skills compared to older cohorts, as well as limited or no job experience (Sackey and Osei, 2006). In addition, young people might not have attained higher levels of education yet, and therefore are less likely to secure formal sector jobs. The information gap between young job seekers and potential employers and barriers to gaining access to financial, physical and social capital for the establishment of businesses are also a major factor of the high incidence of youth unemployment. According to Poku-Boansi and Afrane (2016), youth unemployment reflects the imbalances between the demand and supply of labor. Moreover, there is an issue of quality of labor supply as many of the job seekers either have no skills or have unwanted skills. This highlights a mismatch between academic education and the formal labor market requirements. Even when a person has the required skills, the lack of information and institutional support are some other obstacles for youth job seekers.

1.2.2 Ghanaian Education System

Figure 1.1 represents the Ghanaian school system. Basic education consists of six years of primary school followed by three years of lower secondary school, called Junior High School (JHS). JHS is the last stage of free and universal basic education for the database sample used in this thesis. [2] The Basic Education Certificate Examination (BECE) takes place at the end of JHS and determines whether or not a student enters the upper secondary cycle known as Senior High School (SHS). Grades range from 1 (highest) to 9 (lowest) with an average of 5, which is the passing grade of the examination.

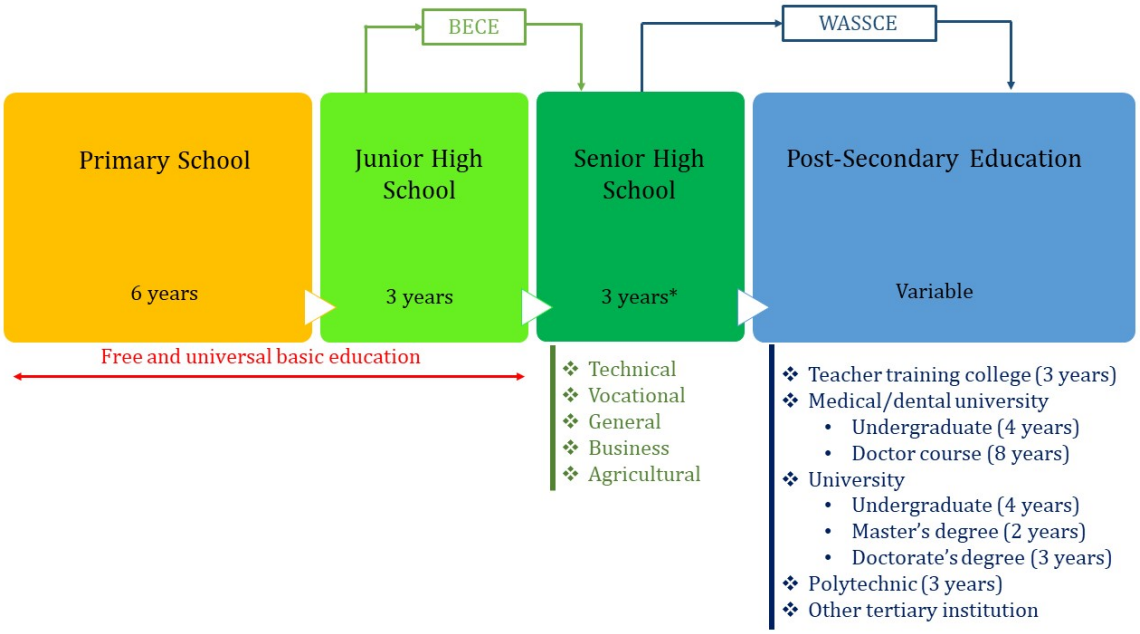
According to Duflo et al. (2017), 70% of JHS students take the BECE and 60% of candidates pass. In addition, about 20% of those admitted do not enrol in the SHS the following year and cost seem to be the main reason (Ajayi, 2014). These different points make access to SHS limited. In fact, transition rate from JHS3 to SHS1 was 50% in 2012 and reached 68% in 2016-17 (Ministry of Education Ghana, 2018b). The three years of SHS [3] end with a standardized examination common to English-speaking West African countries, the West African Senior School Certificate Examination (WASSCE). This exam tests students in four main subjects, namely Mathematics, English, Science and Social Sciences, as well as four other elective subjects. WASSCE result determines admission to higher education.

According to Ministry of Education Ghana (2018b), access to basic education has improved consi-

2. In 2017, the Ghanaian government extended free education to SHS but this does not apply to our sample, who officially completed SHS in 2014.

3. For students entering SHS between 2007 and 2009, there were four years of SHS.

FIGURE 1.1 Ghanaian Education System



Source : EP-Nuffic (2015) and Ministry of Education Ghana.

Note : *Except students entering SHS between 2007 and 2009 for which SHS was 4 years.

derably in recent decades. The gross enrollment ratio (GER) was 104.8% in 2018 according to World Bank (2019) and reaches 85% in JHS. Net rates are generally lower, suggesting that a certain proportion of children enrolled in school are not of the appropriate age. In addition, dropping out of JHS remains a significant problem, with a JHS completion rate of only 66% in 2012, versus 94% in primary school.

At the upper secondary level, the 2010 census reported that only 15% of the population had attended SHS. In 2017, the World Bank estimated that the GER was 30% at SHS and 5% at higher education. Ghana's Ministry of Education figures are a little more optimistic. These show that the GER at SHS has increased from 37% in 2011-12 to 50% in 2016-17 (Ministry of Education Ghana, 2018b). For post-secondary education, GER was estimated to be 14% in 2014-15 (ibid.). Though second cycle education in Ghana is not free, students in public SHS are eligible for subsidies. According to Oberg (2012), the subsidy was 64.35 GHC per student in 2011 and paid directly to schools. In contrast to the lower levels of education, the top public SHS are amongst the most popular schools in the country and attract students who have attended private basic schools and can afford to pay full fees (Oberg, 2012).

According to Ministry of Education Ghana (2018b), inequalities in education persist by income, region and gender. For example, Ghana Statistical Service et al. (2015) estimated that 26.2% of women and 17.5% of men had never received an education in 2014. Furthermore, women are more

likely to drop out of primary school and JHS than men (Ghana Statistical Service, 2012). In 2014, the median number of completed years of education was 4.4 for women and 5.9 for men (Ghana Statistical Service et al., 2015). There are also gender differences in higher education. In 2014-15, only 37% of students applying to public universities were women. In addition, women had a lower probability to be admitted in public university after applying than men (25% for women versus 29% for men).

Ministry of Education Ghana (2018b) highlights that expenditures per student at school level vary significantly by region and are strongly correlated with the level of regional poverty. Of the 450 000 out-of-school children, most of them come from the poorest households and live in the three northern regions (Upper East, Upper West and Northern). Children from the lowest income quintile are 0.37 times as likely to complete JHS compared to students from the highest income quintile. In addition, children from the 20% poorest households are six times less likely to access SHS than others.

According to Oberg (2012), the issues of quality in primary education stimulate demand for private schools. Moreover, there has been an increasing number of private schools in rural areas where there is not sufficient government provision of schools and a shortage of classrooms. Between 1987 and 1998, several private SHSs (around 200) were absorbed into the public system (Oberg, 2012). These schools were designed to raise SHS attendance in areas traditionally under-served by the public secondary education. However, they have struggled to attract sufficient students, as the traditional boarding SHS remaining more popular.

Oberg (2012) counts 505 public and 242 private SHSs in Ghana in 2012. She reports that 91% of the secondary students were enrolled in the public sector, which is much higher than the 77.8% reported by Ghana Statistical Service (2014a) for the same year. In 2018, Ministry of Education Ghana (2018c) estimated that the private sector enrollment (in SHS) was at 6% of total enrollment, while approximately one-third of all SHSs are private schools. What seems to be unanimous however, is that public enrollment at the secondary level is higher than at the primary level (Oberg, 2012). Moreover, urban population is more likely to attend private schools (26.1%) than rural population (12%) (Ghana Statistical Service, 2014a).

Even if public SHSs represent the large part of the SHSs of the country, the choice of focusing on public SHSs has an impact on the population we focus on. As private SHSs are more expensive, the consequence is that they might attract students from richer households (Duflo et al., 2017). In addition, it appears that enrollment in private SHSs is higher in urban areas than rural areas which re-enforces the probability of touching richer households in private schools. However, the public SHSs seem to reach a better level than the private ones (Oberg, 2012). Therefore, by focusing on the public schools, we may miss a part of the population attracted by private SHSs which have a

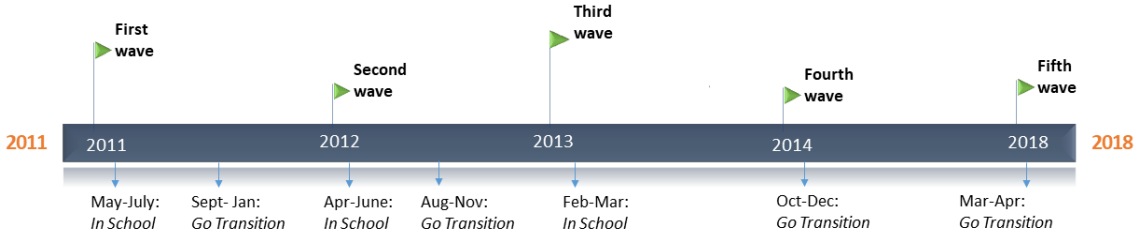
higher probability to be rich, from urban areas, with low results and capacities.

1.2.3 The Panel Survey

The Ghana Opportunities for Transitioning Senior High School Students project was initiated in 2010 by Kim Lehrer and Christopher Ksoll during their postdoctoral studies at the Centre for the Study of African Economies at the University of Oxford. The survey aims to examine young education and employment outcomes, and to follow for several years these Ghanaians. Seven cohorts of students and former students enrolled (and previously enrolled) in SHS have been interviewed across five waves of survey.

As our sample includes both students and former students, we created two types of questionnaires to account for this specificity. The 'In school' questionnaire interviews students in SHS, asking questions about their education and personal information. The 'Go Transition' questionnaire interviews former SHS students about the national exam at the end of SHS (the WASSCE), apprenticeships, post-secondary education and professional activity. Figure 1.2 summarizes the timeline of the different waves of surveys.

FIGURE 1.2 Timeline of Surveys



Anne Duplantier (2021)

Table 1.1 shows the cohorts having been interviewed in each survey, according to the wave and the type of questionnaire. During the first wave of the survey, in 2011, six cohorts were surveyed : three cohorts with the 'In School' questionnaire (cohorts starting SHS between 2008 and 2010), and three cohorts with the 'Go Transition' questionnaire (cohorts starting SHS between 2005 and 2007). The 2011 cohort was not interviewed at that time, as this cohort had not started SHS yet. [4]

During the second wave, in spring 2012, the seventh cohort entered in the panel survey and was interviewed using the 'In School' questionnaire. In 2014, all cohorts should have graduated from

4. The students of this cohort started SHS in September 2011 whereas the survey was conducted between May and July 2011.

SHS, [5] therefore the 'Go Transition' questionnaire was the only one conducted. In 2018, the survey covered five cohorts only. As the cohorts starting SHS in 2005 and 2006 were particularly difficult to reach, we decided to drop these two cohorts from the sample.

TABLEAU 1.1 Waves and Cohorts of the Database

Wave	1		2		3	4	5
Year	2011		2012		2013	2014	2018
Type	In School	Go Transition	In School	Go Transition	In School	Go Transition	Go Transition
	C_{2008}	C_{2005}	C_{2009}	C_{2005}	C_{2009}	C_{2005}	C_{2007}
	C_{2009}	C_{2006}	C_{2010}	C_{2006}	C_{2010}	C_{2006}	C_{2008}
Interviewed	C_{2010}	C_{2007}	C_{2011}	C_{2007}	C_{2011}	C_{2007}	C_{2009}
Cohorts				C_{2008}		C_{2008}	C_{2010}
						C_{2009}	C_{2011}
						C_{2010}	
						C_{2011}	

Note : The school year starts in September and ends in June. Cohort's year is the year a cohort started SHS. For example, the cohort 2005 started SHS in 2005. Anne Duplantier (2021)

The sampling is at school level : 136 public senior high schools were randomly selected amongst the 505 public senior high schools in Ghana. [6] The sample frame used to select the schools is the list of all the public senior high schools in Ghana. Therefore, the universe of the survey is all the students who attend a public SHS in Ghana. The random sample has two levels of weighting : at the regions level, by the number of SHS in 2011 ; then at the schools level, by the number of students. In each selected SHS, one class per level, and, eight students in the selected class, were randomly chosen. In addition, one class of students having theoretically graduated in 2008 and one in 2009 (cohorts 2005 and 2006) were randomly sampled from these schools with eight students per sampled class. The same number of students (56 in total) was selected in each SHS. Therefore the size of the sample is 7616 students. [7] This sample is representative of the public SHS students in each region of Ghana. [8]

5. We use the term "graduates" to refer to both SHS graduates and the few respondents who did not successfully complete SHS. In 2018, they were between 0.10% and 0.94% depending on the cohort).

6. It represents 27% of the whole public SHSs.

7. There are 8 students from 7 cohorts in 136 schools : $8 \times 7 \times 136 = 7616$ individuals.

8. Notice that the number of SHS has increased during the time of the analysis (2011-2018). In 2012, Oberg (2012) reported 515 public and 242 private SHS, while in 2016/2017 the total number of SHS stood at 927 (Ministry of Education Ghana, 2018a). In addition, 124 new SHS were planned, with 44 already completed in 2018. However, according to Ministry of Education Ghana (2018a) private schools made up about half of SHS created between 2011 and 2017.

1.3 METHODOLOGICAL DISCUSSION

1.3.1 The Fieldwork Methods

During the first waves in 2011, 2012 and 2013, the interviews were in-person, either in the SHSs for the 'In School' questionnaire, or at the respondent's place of residence for the 'Go Transition' questionnaire. From 2014, facing financial constraints, we conducted interviews by phone, which is more cost-efficient than in-person interviews, as demonstrated by Sturges and Hanrahan (2004) and Shuy (2003). Indeed, phone interviews allow significant time and monetary savings as enumerators do not have to travel. In cases where a respondent is unavailable, one can call relatives (in order to find the respondent) or call the respondent at another time - this reduces the risk of wasted journey or of losing the respondent completely, as one might in a regular face-to-face interview. The phone survey offers a higher flexibility as the enumerator can easily call another respondent while waiting without losing time. In contrast, in a face-to-face survey, the enumerator loses a lot of time if the respondent is not at home.

Beside the multiple cost-saving feature, there are multiple advantages of telephone interviews. Sturges and Hanrahan (2004) explain that for sensitive topics, respondents may prefer the relative anonymity of telephone. Introverted respondents may feel more at ease on the phone. Responses may be more sincere and honest. According to Shuy (2003), phone interviews also reduce interviewer effects, i.e. unintended influence from the interviewer such as the reactions that may sway the answer of the respondent. For Sturges and Hanrahan (2004), phone interviews allow access to hard-to-reach respondents. This method may give the opportunity to interview respondents who were at first reluctant to participate in a face-to-face interview or who are difficult to access.

Moreover, surveying by phone may provide respondents with more opportunities to participate compared to in-person interviews, allowing us to reach more of those who may not otherwise have had time. Finally, Hoogeveen et al. (2014) recommend that the baseline survey be conducted in person and the subsequent interviews by mobile phone. This allows the first contact with the sampled respondents to be established in person, while still facilitating the time and money of the telephone interviews. Our survey is precisely in this situation, as the three first waves were face-to-face interviews and the two subsequent waves were by telephone.

However, this choice of survey method has some disadvantages. Indeed, the quality of responses, for instance in terms of non-responses to items, is often considered as better using the face-to-face interview method (Shuy, 2003). But Rogers (1976) finds that even though one may think that the presence of the interviewer at the respondent's home would help to convince him/her of the importance of the interview, the quality of data obtained by telephone is comparable to face-to-face interview. According to Hoogeveen et al. (2014), surveying by mobile phone does not limit the

type of questions one can ask. However, the authors recommend that the length of an interview by phone does not exceed 20 to 30 minutes.

Another limitation is that it presupposes that the group of interest owns a telephone. Phone interviews were not typically widespread in Africa because of the low rates of phone ownership. But Hoogeveen et al. (2014) explain that the rapid rise of mobile telephony in Africa changed the situation. For example, mobile phone ownership increased from 9% to 61% between 2004 and 2012 in Tanzania (Hoogeveen et al., 2014). Moreover, our sample is appropriate for a phone survey as educated youth have high probability of owning a mobile phone.

Even if phone interviews offer anonymity to respondents, in-phone survey could be an obstacle when interviewing women. Indeed, in many places with conservative or traditional gender norms, it could be more difficult for girls to isolate themselves and have some privacy to speak freely compared to in-person. In addition, the surveyor cannot ensure that girls are able to speak and answer to questions privately. If such dynamics happen, it could have an impact on the accuracy of the data with a higher bias for women than men. Our survey does not include any questions that capture the ability of the respondent to answer privately to questions. Therefore, we cannot exclude the possibility that the accuracy of the data is worse for girls than for boys. It could conduct to measurement errors stronger for girls and therefore could bias the gender analyses in this thesis.

In cases where neither the respondent nor their contacts [9] answer the phone, the solutions to finding the respondent are limited. If we had chosen face-to-face interviews, we could have walked around the neighborhood to search for and enquire about the person. One risk, then, is losing some respondents. Another problem is the fact that individuals do not conserve the same phone number across time. In this situation, the probability of joining respondents by phone is lower. These obstacles can affect the results if attrition - the loss of sampled respondents over time in a panel database - is not random. The question of attrition will be tackled in the next section, as well as in the Chapter 4 of the thesis which addresses in detail the issue and the way to address it.

1.3.2 Evolution of the Sample and Limiting Attrition

The move of the respondents out of the scope of the survey is a potential cause of attrition. In this case, the survey methodology can strongly affect attrition. Indeed, researchers have to decide whether the surveyors will follow and track sampled movers or not. According to Vaillant (2013), until recently, a lot of surveys do not track respondents when they move inside or outside of the country. Vaillant (2013) explains that the Living Standard Measurement Survey program from the World Bank suggests a dwelling-based follow-up rule. In this case, the sampling unit is the dwel-

9. We collected number phone of family members or friends of the respondents in order to reach them.

ling (instead of the household) and whoever lives there will be interviewed. If a household migrates, this household will be replaced by the new one living in the same dwelling. This methodology eliminates attrition because of migration. But this type of follow-up has some inconveniences. First, migrants are automatically excluded from the sample, so this type of survey does not account for spatial mobility. Thomas et al. (2001) argue that the replacement of sampled households creates significant biases which occur when the mobility is not random. For instance, the results of estimations from that kind of sample can be biased if the move of the household out of the scope of the survey is linked to the subject being studied such as gentrification. Thomas et al. (2001) encourage a tracking survey that follows movers in order to limit attrition due to geographical mobility.

In addition to tracking movers, several methods for reducing attrition rate are used in the Young Lives Study (Outes-Leon and Dercon, 2009). For instance, researchers collect basic information between the waves in order to reach the respondents more easily. They also try to work with the same enumerators and to assign them the same respondents interviewed in the previous waves. They build a follow-up protocol (a guide explaining the steps to follow in order to find the respondent) to be able to track respondents efficiently. Finally, they train enumerators to manage respondents who are not willing to participate in order to avoid, as far as possible, refusals. Winkel and Withers (2000) also suggest some fieldwork procedures to minimize attrition. As in the Young Lives Study, they recommend keeping the same enumerator with the same respondents in each wave. The second idea is to give a gift to the respondents in each wave to encourage individuals to continue to participate. Finally, Zabel (1998) confirms these propositions and finds that the longer the interview and the more changes to the enumerator between waves, the higher is the attrition. He recommends setting up shorter interviews and keeping the same interviewer in order to decrease attrition.

In our survey, we adopted methodologies similar to the ones used in the literature to reduce attrition. We attempt to track all respondents within the country. To do so, we asked enumerators of the 2011 in-person 'Go Transition' survey (outside of the schools) to collect certain information to be able to find respondents in the field for the 2012 in-person survey. We asked for the name of the street, the number of the house, a nearby landmark (such as a church, mosque or shop) and a description of the house to make it easier to find the respondent. The respondent was also asked whether they thought they would be living at the same location in one year, and if not, to where they were planning to move. We then collected personal information such as phone number, email address and Facebook name in order to contact the respondent before the interview. The phone numbers of parents and siblings of the respondents and of three close friends were also collected in case the respondent was not reachable on their own number. Finally, enumerators collected the name and the location of the church or the mosque the respondent regularly attended in order to be able to visit if the respondent could not be located.

Once in the field or on the phone, enumerators used this information to reach the respondent and to confirm that the correct respondent is interviewed. We created a procedure with several steps to facilitate the tracking of the respondents by phone or in-person. As recommended previously, we tried, as far as possible, to keep the same enumerator team and to assign enumerators to the respondents they had already interviewed. We believed that respondents would be more willing to participate if they recognized and trusted the enumerator.

We also gave a small gift [10] to each respondent in order to encourage participation. According to Laurie and Lynn (2008), it is quite usual to give monetary or non-monetary incentives to the respondents of surveys. They list several reasons for using incentives in the form of a gift or money. The first is to encourage respondents to participate or thank them for their participation. Another is to reduce the “non-response bias" by retaining people who are less likely to take part (respondents who are under-represented such as low-income groups). A third reason is to heighten the quality of the responses by giving to respondents the feeling that their participation is important, and giving to enumerators the confidence and the legitimacy to approach the respondent. Moreover, incentives are particularly used with panel surveys in order to encourage regular participation over time. Several articles in the literature analyse the use and the effects of incentives in surveys (Singer et al. (1999), Singer et al. (2002), Ryu et al. (2005), Laurie and Lynn (2008)). In general, they find that incentives increase response rates and improve sample composition. There is little evidence of a significant impact on data quality. Lastly, incentives are recognized as a part of the strategy to reduce attrition in the panel surveys.

Finally, we trained enumerators in how to limit the number of refusals by the respondent. As a last resort, an experienced enumerator was dedicated to the difficult cases of tracking or refusals.

1.3.3 The Questionnaire

The questionnaire for the surveys conducted by phone was shortened. Indeed, a respondent might accept to take one hour to respond to questions face-to-face, but may find this too long if the interview is conducted by phone. The literature suggests that the acceptable length for a phone interview is 20 to 30 minutes (Hoogeveen et al., 2014). Beyond this time, respondents are likely less willing to participate or are less responsive.

The first section of the questionnaire, allows the enumerator to select the right respondent, read the consent form and verify that the respondent agrees to participate. The next section is dedicated to general information about the respondent, and the contact section allows us to collect contact details of the respondents, relatives and friends in order to re-contact the respondent for a future

10. The gift was a bundle to recharge their phone line.

wave of the survey. The fourth section details the SHS the respondent is currently attending (if she is still attending SHS) and her secondary education characteristics. Another section is about the WASSCE tests the respondent has taken and both the expected and actual grades obtained. The post-secondary applications and post-secondary education sections detail the applications the respondent made and the tertiary institutions the respondent attended (or is currently attending).

In the next sections of the questionnaire, we ask questions about apprenticeships and jobs the respondent has done since they left SHS. We collect information on the primary activity, the number of hours spent and the amount earned for each year since the last interview. If the respondent is searching for a job, we ask which methods were used and whether they succeeded or not. The migration section is dedicated to the internal migration history of the respondent, and the last section is dedicated to their well-being, with questions about the satisfaction on several topics including their professional life, financial situation and life in general.

We capture earnings with the following question : “What are your average weekly earnings or take-home profit ?". The advantage of this measure is that it is a quick question, easy to ask. Therefore, it does not burden the survey. However, this methodological choice may have some limits. First, we cannot be sure that this measure captures all the earnings as we do not ask precisely for in-kind incomes, second job earnings, earnings from agriculture work. Moreover, we do not decompose the earnings question for the self-employed earnings (benefits, charges, ...) as it is done in some surveys (for example the Living Standard Measurement Survey (LSMS) from the World Bank).

Second, this measure might allow some approximations as we ask for an average of weekly earnings. Asking for the respondent to compute the average of her earnings results into high probability of measurement error and misreporting. The LSMS survey usually ask the earning for the last month of the previous week. It refers to a more specific time and avoids the respondent computing by herself the average. Asking for the weekly earnings may also lead to measurement error if the worker is paid at another frequency and has to convert her earnings to a weekly frequency. In the LSMS surveys, earnings are asked the way the respondent usually receive them (without specifying a frequency) and then the respondent is asked about the unit of time. Finally, there is no mention of taxes in our measure of earnings. We assume that respondent gives the available earnings (after taxes) but we cannot be sure that is what they understand. Therefore, we cannot be sure of the very high accuracy of the earnings we get from the respondents.

1.3.4 The Field Experience

In September 2014, I prepared the field by modifying and building the questionnaire using the software SurveyCTO. We travelled to Ghana early October. The first two weeks in Ghana were dedicated to the preparation of the fieldwork. We visited the office that would accommodate the

survey team, and searched for and bought phones and bundles that would be used during the survey. We dedicated a large period of time to recruiting and training 15 enumerators, as well as testing the questionnaire and making final corrections. The two following months were devoted to conducting the interviews, managing the team, checking the quality of the data and correcting the mistakes or inconsistencies. Once back in Canada, I continued the corrections and modifications to the database as well as merging the survey waves in order to create the panel database.

In January 2018, I participated in the organization of a new wave of data collection with Kim Lehrer and Bintou Ouedraogo, a master student of the economic department of the Université de Sherbrooke. We mainly re-used the materials we had from the last wave in 2014, but a few tasks were needed before the field : sharing with Bintou the knowledge about modifying the questionnaire on SurveyCTO, building and updating the fieldwork documents, and creating the database needed for the look-up. [11] We arrived in Ghana early March and devoted our time to training and selecting the enumerators, making the last adjustments to the questionnaire, and organizing the last details of the day-to-day field. Kim and I left Ghana two weeks later and left Bintou in charge of the data collection for the next two months. Returning to Canada, I followed-up the survey with Bintou, advised her and answered her questions. Once the corrections of the database were finalized, the last step was to add this fifth wave to the existing panel.

1.4 SAMPLE DESCRIPTION

This section describes the characteristics of the sample by focusing on education and labor market outcomes. As explained previously, educated youth are crucial for the economic development of a country. Hence, providing stylized facts about both the secondary and post-secondary education process (results, attendance, costs, access to post-secondary education, expectations) will help policymakers to improve the education system. It is also crucial to understand how these graduates from the SHS or the post-secondary education enter the labor market and on which conditions. Knowing the characteristics of the young Ghanaian searching for a job will help to understand youth unemployment. The first part focuses on the primary individual and household characteristics of the respondents (mainly characteristics allowing to assess the poverty situation of the respondents). I then describe the secondary education situation and how the respondents access to post-secondary education. Finally, the situation of the youth that entered into the labor market is presented.

11. The look-up is information from previous waves that we use in the questionnaire of the current wave of survey. For example, by knowing that a respondent already graduated from SHS in 2014, we will not re-ask the question in 2018. Therefore, we extract the information available from the previous waves for adapting the questionnaire according the situation of the respondent.

1.4.1 Primary Individual and Household Characteristics

Primary individual and household characteristics of the respondents are presented in the following tables and figures. The primary individual and household information represents the initial information about the respondent. The questions for these characteristics have been asked only once for each respondent (the first time the respondent was interviewed) and are considered here as invariant. Thereby, we do not use the panel dimension and keep one observation per respondent. As 1048 individuals have never been interviewed, [12] the sample size is lower than the theoretical sample (6568 versus 7616). As it can be seen in Table 1.2, 60% of the respondents are male, meaning that the men have an easier access to the SHS than the women. This is quite close to the official number from Ghana Statistical Service (2012) that finds that 55.05% of SHS graduates in 2010 were male. This is also consistent with Duflo et al. (2017) who report that in 2010, girls in Ghana were 20% less likely to reach SHS.

Half of the respondents were born in a rural area, which is only slightly higher than the 49.1% found by the 2010 census (Ghana Statistical Service, 2012). This difference may be due to differences in rural residence for our sub-sample compared to the Ghanaian population, differences in rural/urban residence between 1990 (when, on average, our sample was born) and 2010, or due to differences in the definition of rural between the 'Go Transition' survey and the census. Indeed, the 2010 census considers a zone to be rural if the population is below 5000 inhabitants (Ghana Statistical Service, 2012). However, the 'Go Transition' survey uses a subjective measure as the respondent is asked whether their mother was living in a rural or urban area when they were born, without providing a precise definition. Two respondents born in the same area can answer this question using different terms (rural or urban). With regard to gender, there are fewer girls than boys born in rural areas (difference statistically significant). Finally, the respondents have on average four siblings and the girls come from smaller households (difference statistically significant). This may suggest that access to education in rural areas and in larger households is more difficult for girls. It is consistent with the report from Ministry of Education Ghana (2018b) highlighting that students from poor households and/or rural areas have five to six times less chance to access SHS.

Figure 1.3 presents the distribution of respondents according to the education level of their parents. On average, the father of the respondent appears to be more educated than their mother. While the mothers without any education are 30% of the sample, the fathers are 19% - more than ten percentage points of difference. For 43% of the respondents, the father has an education level equal or superior to them (i.e. at least SHS), compared to 22% respondents' mother. Looking at the post-secondary education, the difference is the most important : 7.8% of the mothers attended a post-secondary institution while this is the case for 20.7% of the fathers. This gender gap in education

12. See Chapter 4 about attrition.

TABLEAU 1.2 Primary Individual Characteristics

	(1) Whole	(2) Girls	(3) Boys	(4) P-value
Male (%)	60.06	-	-	-
Born in Rural Area (%)	51.95	48.16	54.49	0.000
Number of Siblings	4.09	3.89	4.22	0.000
	(2.36)	(2.20)	(2.45)	
Number of Observations	6568	2620	3940	

Note : One observation per respondent. Standard deviations are in parentheses. Because 1048 individuals have never been interviewed, the sample size is lower than the theoretical sample (7616). Column (4) is the p-value of the t-test difference between (2) and (3). Anne Duplantier (2021)

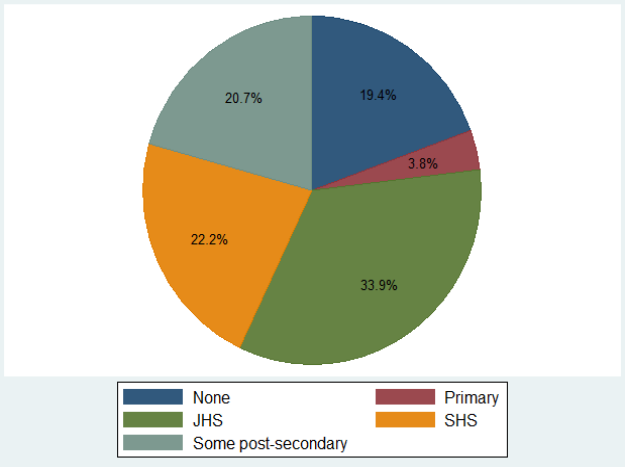
of parents is consistent with the report from Ghana Statistical Service et al. (2015) that found that the median number of completed years of education is 4.4 for women and 5.9 for men.

Figure 1.4 presents the main occupation of the parents during the majority of their life. As the question about the occupation was asked only in the 2011 and 2012 'Go Transition' waves, the number of observations is smaller. [13] The main occupation of the parents is very different according to the gender of the parent. The majority (59%) of mothers are self-employed in the trading sector. Another important sector where mothers work is agriculture. But very few of them (less than 5% for each type) are government employee, teacher or self-employed in the services. Concerning fathers, the most common first occupation is in the agriculture sector (32%), followed by government employment (18%), self-employment in services (15%) and other waged employment (14%). In general, 65.7% of mothers are self-employed (of all types : services, trading, manufacturing) versus 29.7% of fathers. This suggests that mothers have higher probability to occupy precarious employment than fathers, as the self-employed (with farmers) are known to be the most unstable and vulnerable workers in developing countries. Indeed, according to Cho et al. (2016), 80% of self-employed in SSA are poor (i.e. living with less than two United States Dollar (USD) per day for per capita household consumption).

Table 1.3 presents the main characteristics of the respondents' households. The questions about the household characteristics were only asked to Form 1 students during 'In School' waves. Consequently, the sub-sample of household characteristics ($n = 1893$) is composed of the 2008 cohort (interviewed during the 2011 'In School' wave) and the 2009 cohort (interviewed during the 2012 'In School' wave). This sub-sample is representative of the youth in secondary schools at the regional level, as each cohort includes students from everywhere. As the only difference between the

13. The respondents not interviewed during these waves and interviewed in 2011 and 2012 with the 'In School' questionnaire never answered this question. Respondents from three over seven cohorts (the 2009, 2010 and 2011 cohorts) are in this situation.

(a) Father



(b) Mother

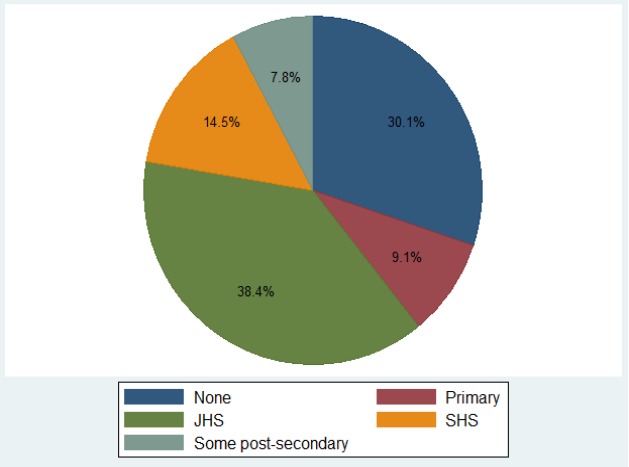
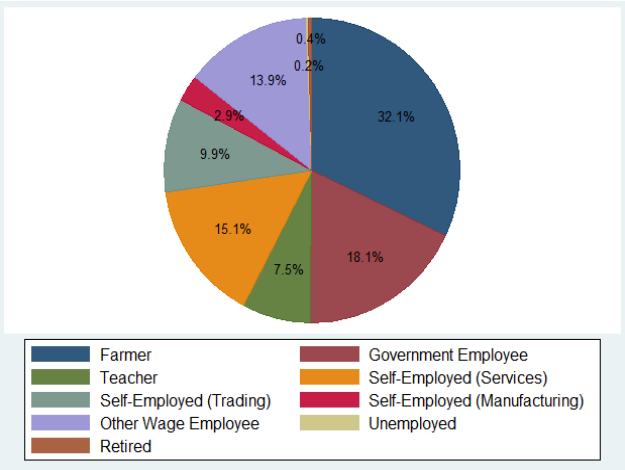


FIGURE 1.3 Distribution of the Respondents According to the Education Level of the Parents

Note : One observation per respondent (n = 5999). Anne Duplantier (2021)

(a) Father



(b) Mother

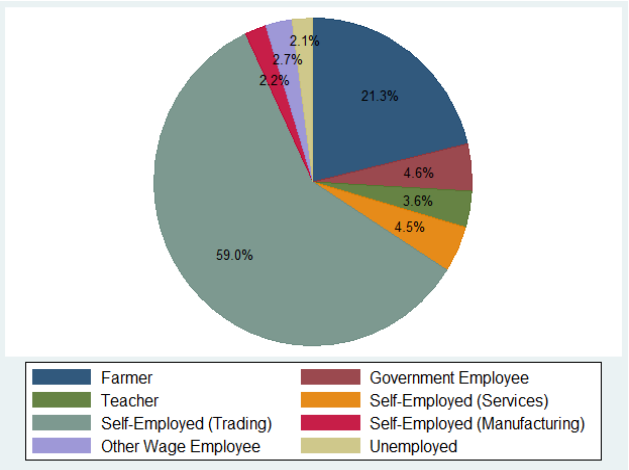


FIGURE 1.4 Distribution of the Respondents According to the Main Occupation of the Parents

Note : One observation per respondent (n = 3203). Anne Duplantier (2021)

cohorts is the age of the respondents, this sub-sample remains representative of the whole sample unless these characteristics change over time.

On average, the respondents come from households with seven members. Consistently with Table 1.2, girls have generally smaller households than boys (difference statistically significant). We asked if all the children between 6 and 11 years old attend school, to which 13% of the respondents answered “No" and 63% “Yes", with no significant difference between girls and boys. About 95% of the respondents have a housing with modern roof [14] and 80% use electricity as main source of lighting. This is quite consistent with the report from Ghana Statistical Service et al. (2015) that found that 78.3% of households have access to electricity. In 63% of the households, the main source of drinking water is indoor or protected, whereas 32% use a borehole and 5% take water from the river or from the rain. [15] For 52% of the respondents, the household owns a working stove and 90% of them own a working iron.

There are some statistically significant differences in the housing characteristics according to gender. On average, girls come from housing with better infrastructures : 96% of girls have a housing with a modern roof (versus 94.5% for boys), 84% use electricity as the main source of lighting (versus 78% of boys). More girls (64% versus 60% for boys) have a safe source of drinking water (indoor or protected source), whereas more boys use a borehole or river/rain water as the main source of drinking water. For 57% of the female respondents, the household owns a working stove (versus 48% for the boys) and 92% a working iron (versus 89% for the boys). In conclusion, it seems that girls attending SHS come from less poor, less rural and smaller households than boys. This is consistent with the literature on education in developing countries finding disparities between girls and boys (Sutherland-Addy, 2008). In general, households prefer to invest in boys' education than girls' education. In consequence, the poorer households are less likely to send their girls to school. Indeed, Sutherland-Addy (2008) highlights that enrollment of girls in secondary school is related to the wealth of the family in SSA countries.

1.4.2 Secondary Education Characteristics

Table 1.4 presents some characteristics of the secondary education of the respondents. In Part A, the statistics are presented for one observation per respondent, while the Part B uses the longitudinal dimension of the database. This means that several observations per respondent are used when

14. 'Modern roof' refers here to corrugated iron sheets, cement/concrete, asbestos/slate, or roofing tiles, versus palm leaves/raffia/thatch, wood, mud bricks/earth, bamboo, or other.

15. For the source of drinking water, the categories are : i) Borehole, well (with pump or not, protected or not), or other; ii) Indoor plumbing, inside standpipe, sachet/bottled water, standpipe/tap (public or private outside) pipe in neighbouring household, water truck/tanker, or water vendor; iii) River/stream, rain water/spring, or dugout/pond/lake/dam.

TABLEAU 1.3 Household Characteristics

	(1) Whole	(2) Girls	(3) Boys	(4) P-value
Household Members	6.98	6.65	7.21	0.001
	(4.16)	(3.32)	(4.65)	
All Children 6-11 Attend School				0.863!
No (%)	12.89	13.14	12.71	
Yes (%)	63.55	63.90	63.30	
No Children (%)	23.56	22.96	23.99	
Modern Construction Material Used for Roof (%)	95.19	96.17	94.50	0.046
Electricity Main Source of Lighting (%)	80.56	83.80	78.27	0.001
Main Source of Drinking Water (%)				0.006!
Borehole (%)	32.12	29.21	34.17	
Indoor or Protected Source (%)	63.23	64.22	60.41	
River or Rain Water (%)	4.65	3.57	5.41	
Own a Working Stove (%)	52.03	57.02	48.51	0.000
Own a Working Iron (%)	90.44	91.84	89.45	0.041
Number of Observations	1893	784	1109	

Note : One observation per respondent. Standard deviations are in parentheses. The questions about the household characteristics were asked to Form 1 students during 'In School' waves. Column (4) is the p-value of the t-test difference between Columns (2) and (3). ! means that the p-value is from the χ^2 test. Anne Duplantier (2021)

available. On average, respondents entered into SHS with an aggregate score of 18.68 at the terminal BECE exam at the end of JHS. [16] There is no real difference in score between girls and boys even if girls have a slightly better average score (18.55 versus 18.77). According to Duflo et al. (2017), pass rates are quite low : 70% of JHS entrants go on to take the BECE and 60% of BECE takers pass. Moreover, Ajayi (2014) highlights that about 20% of those admitted do not enroll in SHS the following year and many cite costs as the reason.

There are eight different programs available for SHS students in Ghana. The eight programs have the same core subjects - English, mathematics, integrated science, social studies - and differ on elective subjects - agriculture, business, general arts, home economics, science, technical, visual arts and vocational. Figure 1.5 displays the distribution of respondents according to the SHS elective subjects they chose. The three main subjects chosen by the male students are Business (25.6%), Science (24.5%) and General Arts (21%), while girls chose Business (24.7%), General Arts (24.4%) and Home Economics (21.5%). The fourth subject chosen by boys is Agriculture, whereas it is Science for the girls. However, this is only representative of the choices of Ghanaian students if the size of the classes is not correlated with the elective subjects. In the case where the

16. Each subject scores from 1 (best) to 9 (lowest) with an average at 5, which is the pass mark for the exam. At the aggregate level, the best score is 5 and the lowest is 45, while the student has to get 30 or lower (i.e. a better score) to enter into a SHS.

size of the classes varies non-randomly with the program, the statistics about the elective subjects would not be perfectly representative.

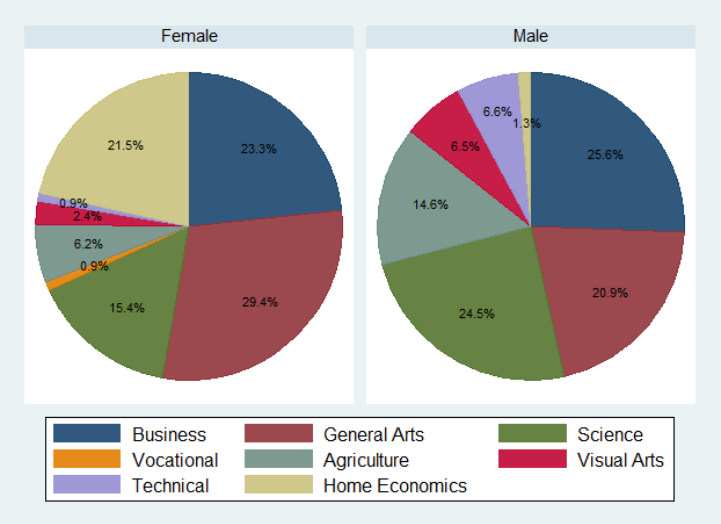
During the 2012 and 2013 'In School' waves, the respondents were asked to estimate the annual fees they paid for SHS. The values have been deflated to allow comparisons and are expressed in base year 2011. The annual fees were on average 498 Ghana cedis (GHS) in 2012 and 565 GHS in 2013 (Table 1.4). It seems very close to what Duflo et al. (2017) report : they estimate that the government-approved tuition fees were 500 GHS in 2011. The authors highlight that this amount represents a very large sum in a country where the annual GDP per capita in 2011 was about 2400 GHS. Figure 1.6 shows the distribution of the respondents according to the person who pays the majority of the SHS fees. For 82.5% of the respondents, their parents are covering their SHS tuition fees. Only 2.7% of the respondents reported that scholarship pays the majority of the SHS fees, while other respondents are helped by relatives (12.4%) or financing education by themselves (1.8%). It worth noticing that the introduction of free SHS in September 2017 is an important reform as the cost of SHS is one of the key barrier to attendance (Ministry of Education Ghana, 2018b). Thus, this measure is expected to allow a better access and equity at the secondary education level.

During the 'In School' waves, questions were asked about the presence at school of both the respondent and their professors. [17] In response, 13% declared that they were absent for at least one lesson during the week before the interview. There is a small difference according to the gender of the respondents : 12.75% of the girls reported that they missed at least one lesson the week before the interview versus 13.33% of the boys (but the difference is not statistically significant). Concerning the teachers, 12% and 18% of the respondents reported an absence for at least one lesson the week before the interview for their Math and English classes, respectively.

Figure 1.7 displays the distribution of the respondents according to the number of WASSCE exams taken. In Ghana, students take the May/June WASSCE at the end of the SHS. Then, should they want to improve their exam score, students can take an additional exam, the Nov/Dec WASSCE, the number of times they want and by choosing the subjects they want to take. On average, 63% of the respondents took the WASSCE exam only once, meaning that they passed only the May/June WASSCE, whereas 30% of the respondents took the Nov/Dec WASSCE once, and around 7% more than twice. With regard to gender, there are statistically significantly more girls than boys who took the Nov/Dec WASSCE (31% versus 30%), meaning that girls have a higher probability to be dissatisfied with their first WASSCE score. In addition, 9% of the female respondents took a WASSCE exam more than twice, compared to 5.4% of the males (statistically significant difference).

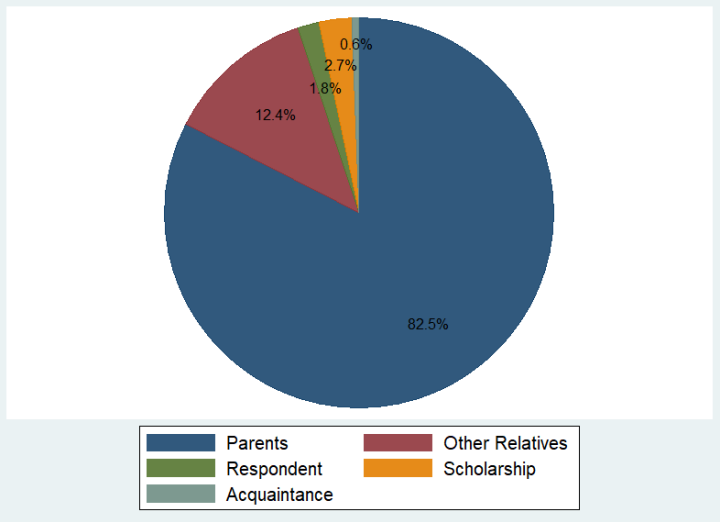
17. Questions are as follow : "How many days in the last seven days were you absent from at least one lesson ?" and "How many days in the last seven days was your Math/English teacher absent from at least one lesson ?"

FIGURE 1.5 Distribution of the Respondents According to the SHS Subject



Note : One observation per respondent (n = 6375). Gender difference's χ^2 test p-value = 0.000. Anne Duplantier (2021)

FIGURE 1.6 Who Pays the Majority of the SHS Fees ?



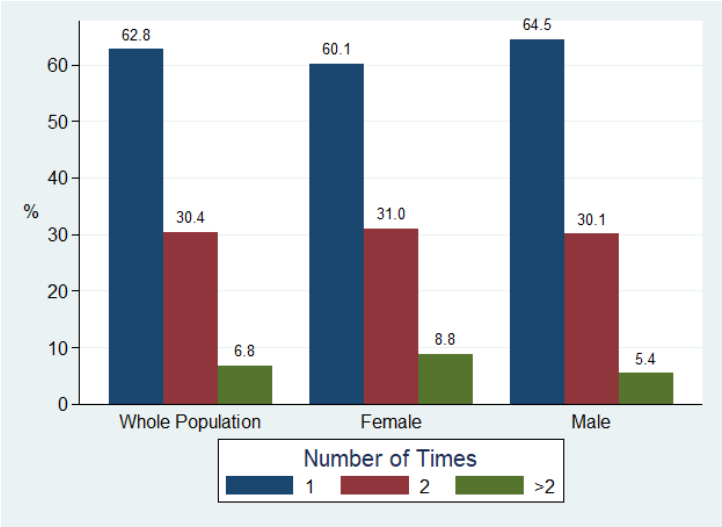
Note : Several observations per respondent (n = 11 980). The question was asked in all the waves, except 2011 'In School' and 2018 'Go Transition'. Anne Duplantier (2021)

TABLEAU 1.4 Secondary Education Characteristics

	(1) Whole	(2) Girls	(3) Boys	(4) P-value
PART A - One Observation per Respondent	6568	2597	3886	
BECE Aggregate Score	18.68	18.55	18.77	0.088
	(6.56)	(6.51)	(6.59)	
Number of Applications	1.43	1.46	1.44	0.778
	(1.45)	(1.43)	(1.46)	
PART B - Several Observations per Respondent				
Aggregate Score of May/June WASSCE [$n = 8584$]	18.16	18.62	17.89	0.000
	(4.33)	(4.19)	(4.39)	
SHS Fees in 2012 [$n = 2532$]	498.06	513.60	486.77	0.028
	(350.75)	(334.15)	(362.03)	
SHS Fees in 2013 [$n = 1693$]	565.05	596.13	542.71	0.007
	(447.55)	(530.30)	(375.72)	
Respondent Absent Last Week [$n = 6954$]				0.238
None Day (%)	86.91	87.25	86.67	
At Least One Day (%)	13.09	12.75	13.33	
Math Teacher Absent Last Week [$n = 6935$]				0.351
None Lesson (%)	88.00	88.18	87.88	
At Least One Lesson (%)	12.00	11.82	12.12	
English Teacher Absent Last Week [$n = 6932$]				0.308
None Lesson (%)	82.07	82.34	81.87	
At Least One Lesson (%)	17.93	17.66	18.13	

Note : Part A uses data with one observation per respondent, while Part B uses data with several observations per respondent. The question about the amount of the SHS fees was asked in the 2012 and 2013 'In School' waves. The values have been deflated and are expressed in base year 2011. The questions about the absence in class were asked during the 'In School' waves. Column (4) is the p-value of the t-test difference between Columns (2) and (3). Standard deviations are in parentheses. Anne Duplantier (2021)

FIGURE 1.7 Distribution of the Respondents According to the Number of WASSCE Exams



Note : One observation per respondent (n = 6568). Gender difference's χ^2 test p-value = 0.000. Anne Duplantier (2021)

Table 1.5 presents the average number of WASSCE by cohorts. For every cohort, the average number of May/June WASSCE is between 0.98 and 0.99 except for the 2008 cohort (0.94). In this cohort, some respondents did not take the May/June WASSCE at least once. There are higher differences between cohorts regarding the Nov/Dec WASSCE. The 2011 cohort took the Nov/Dec WASSCE only 0.36 times, while the 2006 cohort has the highest average number with 0.57. In Part B of table 1.4, we present the average May/June WASSCE aggregate score for the whole population and by gender. The average aggregate score for the May/June WASSCE is 18.16 (with 6 the best score, 54 the lowest, and 24 the minimum score allowing to enter into university [18]). Girls have on average a lower score (18.62) than boys (17.89) and the difference is statistically significant.

TABLEAU 1.5 Average Number of Exams by Cohort

	C_{2005}	C_{2006}	C_{2007}	C_{2008}	C_{2009}	C_{2010}	C_{2011}
May/June WASSCE	0.998	0.997	0.998	0.941	0.989	0.983	0.982
	(0.04)	(0.05)	(0.04)	(0.23)	(0.10)	(0.13)	(0.13)
Nov/Dec WASSCE	0.54	0.57	0.52	0.41	0.44	0.43	0.36
	(0.73)	(0.72)	(0.68)	(0.61)	(0.61)	(0.65)	(0.55)
All WASSCE	1.54	1.57	1.51	1.35	1.43	1.41	1.34
	(0.73)	(0.72)	(0.68)	(0.59)	(0.60)	(0.65)	(0.55)
Number of Observations	661	764	963	822	841	809	798

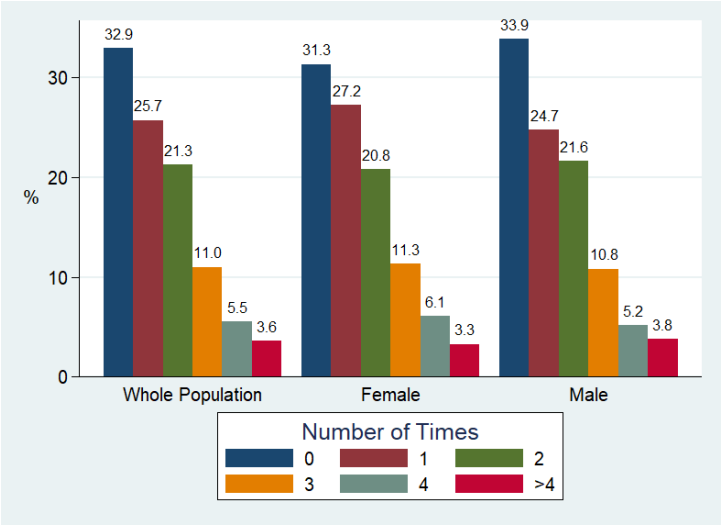
Note : One observation per respondent. Standard deviations are in parentheses. Anne Duplantier (2021)

1.4.3 Applications Characteristics

After SHS, students can apply to a post-secondary institution to pursue their education. In Ghana, there are several types of tertiary institution : university, polytechnic college, teacher training, nursing, technical/vocational college. The characteristics about the applications made by the senior high school graduates are presented in Tables 1.6 and 1.7 as well as in Figures 1.8 to 1.13. On average the SHS students applied to 1.43 post-secondary institutions (cf Part A of Table 1.4). There is no statistically significant difference between girls and boys. Figure 1.8 displays the distribution of respondents according to the number of institutions they applied to. Respondents who never applied to post-secondary institution represent 33% of the sample, while 26% applied one time to a post-secondary institution and 21% applied two times. Regarding by gender, girls applied more than boys, as male respondents are 34% having never applied versus 31% of girls. This is the case for all categories of number of applications and the differences are statistically significant.

18. However, for a lot of programs at university, the cutoffs are higher : 7 for a Bachelor of Medicine or a Bachelor of Laws, 8 for a Bachelor of Pharmacy or a B.Sc. in Administration, 14 for a B.Sc. in Engineering Sciences (cf `https://admission.ug.edu.gh/applying/content/cut-points-2018-19`).

FIGURE 1.8 Distribution of the Respondents According to the Number of Applications Made



Note : One observation per respondent (n = 6568). Gender difference's χ^2 test p-value = 0.052. Anne Duplantier (2021)

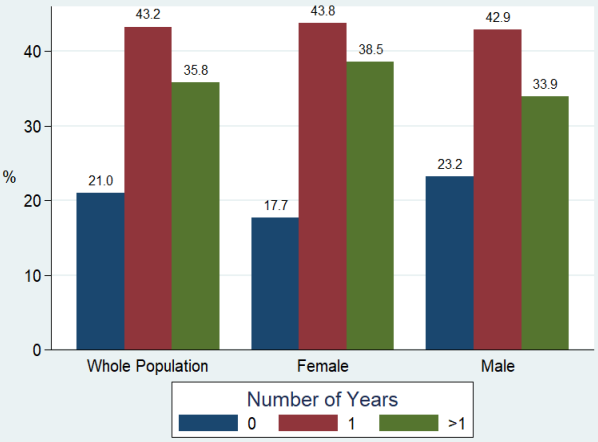
In the following application statistics, we will focus only on the first application made by the respondent. Indeed, the average number of applications is 1.43 per student. Moreover, the first application is usually made just at the end of SHS and reflects the best what students can claim and achieve directly after SHS.

The first time that students apply to post-secondary institution, they do it on average 1.46 years after graduation (cf Table 1.7). Girls apply later than boys (1.56 years versus 1.39 years for the first application) and the difference is statistically significant. Figure 1.9 (a) shows the distribution of the respondents according to the number of years between the graduation and the first application. On average, 21% of the respondents apply for the first time the same year they graduate, 43% one year after, and 36% more than one year after graduation. The figure confirms that girls apply later than boys (difference statistically significant). Only 18% of girls apply the same year of graduation (versus 23% for boys). But 38.5% of the girls apply more than one year after graduation (versus 34% of the boys).

In Figure 1.9 (b), the duration between the graduation and the first application is broken down by cohort. There are important variations according to the cohorts. Respondents from the 2011 cohort take the most time before applying to the first post-secondary institution (52% of them apply more than a year after graduation). By contrast, 30% of the respondents from the 2008 cohort apply to the first post-secondary institution the same year of their graduation, and only 28% of them applied more than a year after.

Figures 1.10 and 1.11 show the distribution of the respondents according to the type of the institution and the degree asked for the first application. As a first application, 40% of the respondents

(a) Whole Population and by Gender



(b) By Cohort

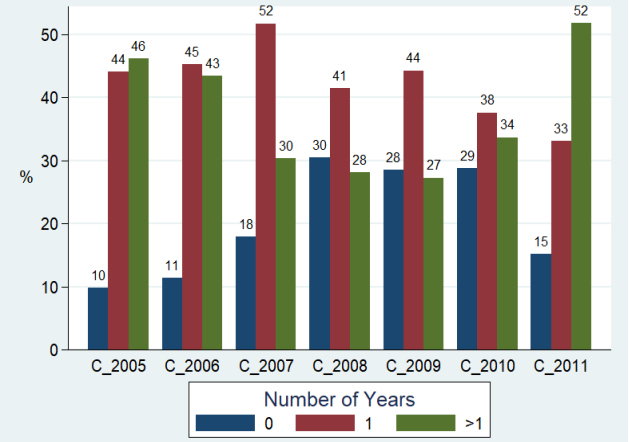


FIGURE 1.9 Time Between Graduation and the First Application

Note : One observation per respondent (n = 4376). Gender difference's χ^2 test p-value = 0.000. Anne Duplantier (2021)

apply to university, 21% to nursing, 18% to teacher training and 11% to a polytechnic college. Looking at gender difference, boys have a higher probability to apply to university than girls (46% versus 30%). Boys apply also slightly more to teacher training (18.5% versus 16.5%) and to polytechnic college (12.5% versus 9%). But 35% of girls apply to nursing studies versus 11.5% for the boys. These gender differences are statistically significant.

Concerning the degree applied, 45% of the respondents apply for a diploma (usually from a technical/vocational college, teacher training or nursing) and 33% for a bachelor degree (from university). A greater proportion of boys apply for a bachelor degree than girls (39.5% versus 24%), who apply in the majority for a diploma (51.5% versus 41% for the boys). Again, there is a higher probability to apply for a certificate (usually from a technical/vocational college, teacher training or nursing) for a girl (17%) than a boy (10%). In general, girls apply for a degree shorter and less valuable than boys. These differences are statistically significant. It is consistent with the choices of tertiary institutions : girls apply more to professional post-secondary institution than boys. This drives to a higher probability for girls of applying to a short degree. Moreover, the percentage of girls applying to a bachelor degree is lower than the proportion of girls applying to university. This means that even when girls apply to university, they do not choose bachelor degrees but shorter degrees such as diplomas and certificates.

Figure 1.12 shows that 71.6% of the respondents only use the May/June WASSCE score for their first application, while 24% of the respondents use a combination of the May/June and the Nov/Dec WASSCE scores. This means that they considered that their May/June aggregate score was not high enough and they improved their initial score with Nov/Dec scores in order to have better chance to

(a) Whole Population (b) By Gender

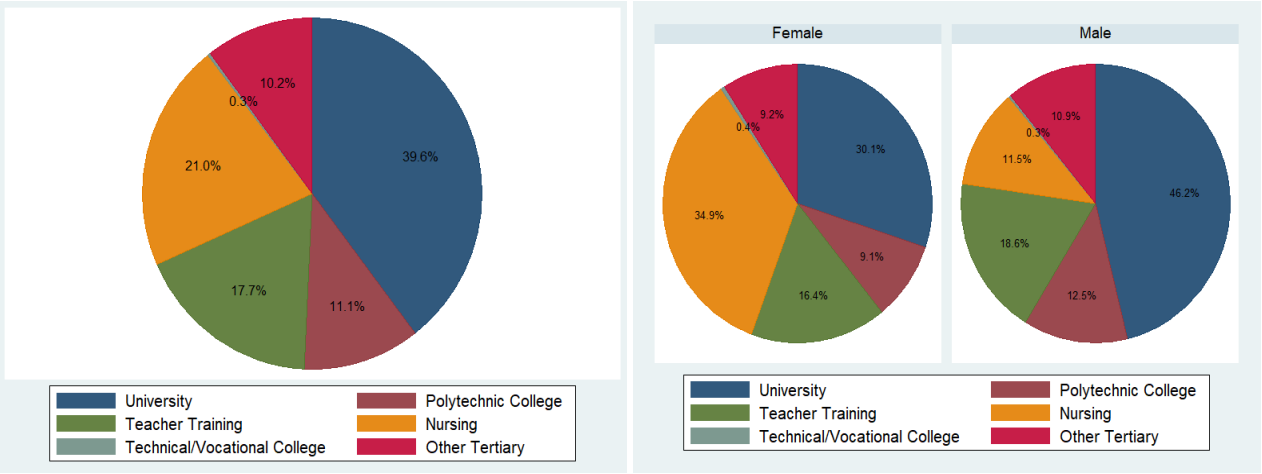


FIGURE 1.10 Distribution of the Respondents According to the Institution Type of the First Application

Note : One observation per respondent (n = 4399). Gender difference's χ^2 test p-value = 0.000. Anne Duplantier (2021)

(a) Whole Population (b) By Gender

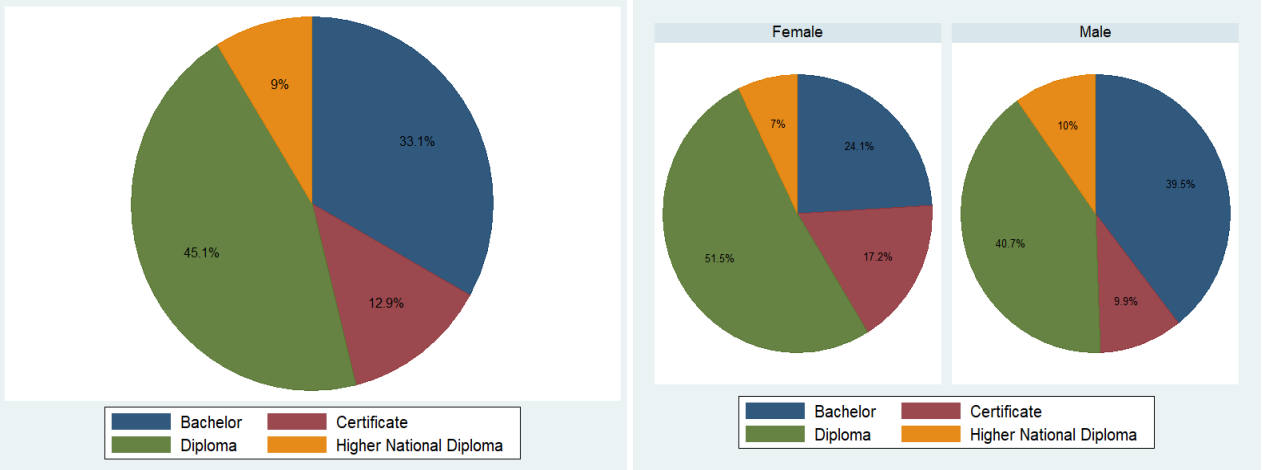


FIGURE 1.11 Distribution of the Respondents According to the Institution Degree of the First Application

Note : One observation per respondent (n = 2816). Question asked during the 2014 and 2018 'Go Transition' waves. Higher national diploma is equivalent of the two first years of a bachelor degree but for polytechnic colleges. Gender difference's χ^2 test p-value = 0.000. Anne Duplantier (2021)

enter into the post-secondary institution they targeted.

Tables 1.6 and 1.7 present the average cost of the first application by gender and by cohort. On average, a respondent spends 73 GHS for the first application. It appears that girls spend a little bit less than boys but the difference is not statistically significant. In Table 1.6, the total cost of all the applications that the respondents made is presented, with the number of applications made and the average cost by application. The values are deflated in order to allow comparison between years and expressed in base year 2011. On average, a graduated student from SHS spends 145.54 GHS for all the post-secondary applications they did. By cohort, the amount spent in applications is decreasing with the age of students : the older students - from the cohort C_{2005} - spend on average 194.37 GHS for all their applications versus 92.28 GHS for the youngest students - from the cohort C_{2011}. It is quite consistent, as the number of all applications made by the respondents is also increasing with the age of respondents : students from C_{2005} cohort made on average 2.78 applications while students from C_{2011} cohort made only 1.65. On average (taking into account the number of applications), the cost of one post-secondary application is 65.72 GHS. Once divided by the number of applications, the differences of costs between cohorts are considerably smaller (from 61.27 GHS for C_{2009} cohort to 70.61 GHS for C_{2005} cohort).

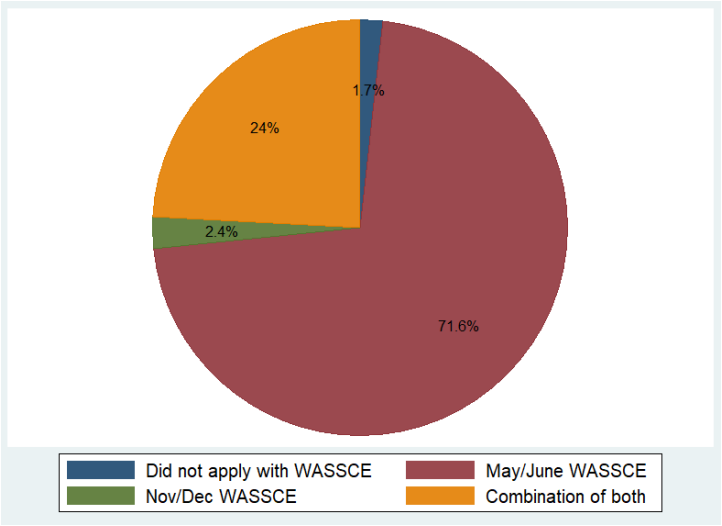
TABLEAU 1.6 Cost of All the Applications

	Whole	By Cohort						
		C_{2005}	C_{2006}	C_{2007}	C_{2008}	C_{2009}	C_{2010}	C_{2011}
Total Cost (All App)	145.54	194.37	166.42	148.07	127.45	113.22	113.79	92.28
	(130.70)	(202.95)	(137.30)	(136.45)	(79.10)	(68.69)	(67.72)	(51.71)
Number of Applications	2.12	2.78	2.52	2.30	1.98	1.82	1.77	1.65
	(1.24)	(1.49)	(1.42)	(1.31)	(1.06)	(0.99)	(0.95)	(0.89)
Average Cost	65.72	70.61	67.95	65.08	65.34	61.27	62.86	62.32
	(55.37)	(75.14)	(59.84)	(75.48)	(23.80)	(27.03)	(25.90)	(30.73)
Number of Observations	4393	531	590	810	756	666	595	445

Note : One observation per respondent. Standard deviations are in parentheses. The total cost is the cost of all the applications made by the respondent. The average cost is the total cost divided by the number of applications made by the respondent. Anne Duplantier (2021)

The last question about the process toward post-secondary education seeks to find out whether the application was successful. For 64% of the respondents, the first application was successful (cf Table 1.7). Girls have a lower probability to have a first successful application than boys (63% versus 65%) but the gender difference is not statistically significant. Amongst students who have a successful application, 83% of them accepted the offer. Girls accepted more frequently the offer from the first successful application than boys (difference statistically significant). Figure 1.13 shows that, amongst those who refused an offer for the first application, 44% did so for financial reasons, and 32% because they had already accepted or paid for another offer. Looking at the figure by gender confirms the evidence that girls come from more well-off households. Indeed, male students have a higher probability (48%) to refuse the offer of their first application for financial is-

FIGURE 1.12 WASSCE Score Used for the First Application



Note : One observation per respondent (n = 3238). Question not asked in 2018 wave. Anne Duplantier (2021)

sues than female students (37%). The gender differences in the reason for refusing the first accepted offer are statistically significant.

TABLEAU 1.7 First Tertiary Application's Characteristics

	(1) Whole	(2) Girls	(3) Boys	(4) P-value
Years btw Graduation and Application [$n = 4376$]	1.46	1.56	1.39	0.000
	(1.27)	(1.31)	(1.25)	
Cost of Application [$n = 3306$]	72.88	71.54	73.78	0.179
	(68.99)	(87.06)	(53.87)	
Application Successful (%) [$n = 6066$]	64.11	62.88	64.93	0.256
Accepted the Offer (if successful) (%) [$n = 3890$]	82.78	85.27	81.15	0.010

Note : One observation per respondent. Standard deviations are in parentheses. The question about the cost of the application was not asked in the 2018 'Go Transition' wave. The values are deflated (base year 2011). Column (4) shows the p-value of the t-test difference between Columns (2) and (3). Anne Duplantier (2021)

1.4.4 Post-Secondary Education Characteristics

In the following tables and figures we describe the post-secondary attendance of the respondents. Figure 1.14 presents the percentage of individuals who are currently (at the time of the survey) attending a post-secondary institution, by cohort and year. Quite consistently, in 2011 and 2012, the oldest students (from C_{2005} and C_{2006} cohorts) have a higher probability to be currently attending a post-secondary institution (between 44% and 52%). Over time, the oldest students leave or complete their studies in post-secondary institution and other students start attending. In 2018,

(a) Whole Population

(b) By Gender

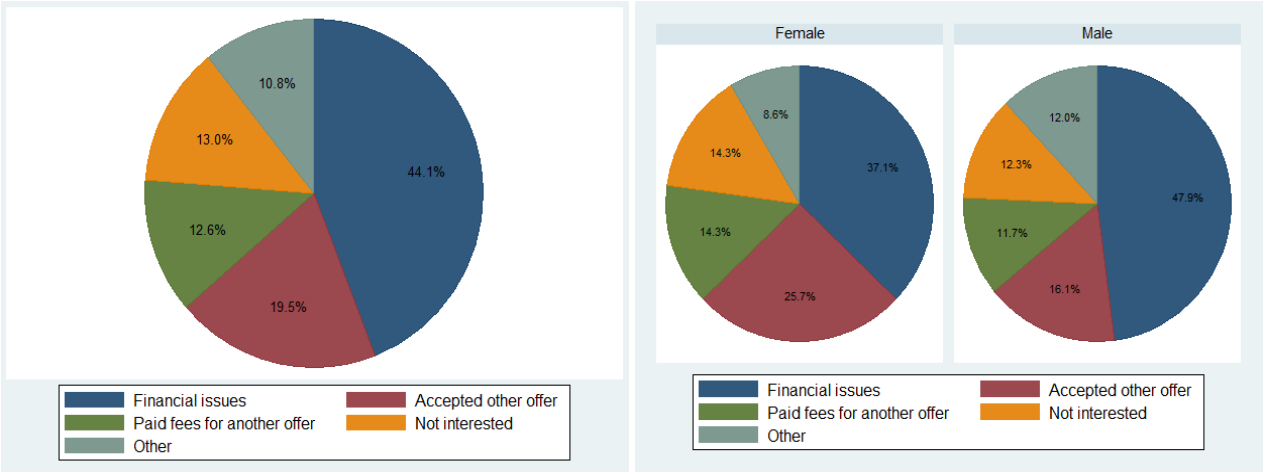


FIGURE 1.13 Reason for Refusing the Offer of the First Application

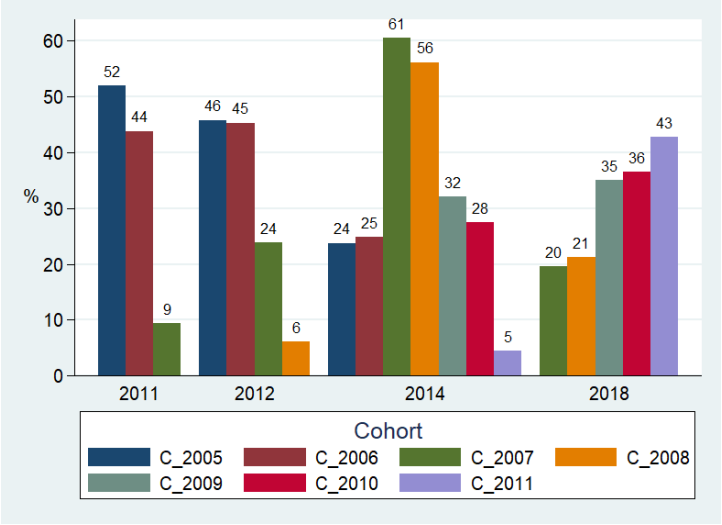
Note : One observation per respondent (n = 392). Gender difference's χ^2 test p-value = 0.033. Anne Duplantier (2021)

43% of the youngest students (from C_{2011} cohort) were attending a post-secondary institution and around 20% of the students from the C_{2007} and C_{2008} cohorts were still attending a tertiary institution, having graduated from SHS in 2011 and 2012. This suggests that people stay for a long time in post-secondary education. Another explanation is that these respondents enter into the post-secondary system very late, which is consistent with the time gap between the graduation and the year of the first application shown previously.

We checked in the data which assumption is the most likely. We first computed the average time between the SHS graduation year and the post-secondary institution starting year. We find that there is on average two years between the year of graduation from SHS and the time when respondents start attending a post-secondary institution. Figure 1.15 shows that 42% of the students from the C_{2007} cohort start attending a post-secondary institution three years after graduating from SHS. And 54% of the students from the C_{2008} cohort start attending a post-secondary institution two years after graduating from SHS. We also computed the average time of post-secondary attendance. On average, respondents attended a post-secondary institution for two years (cf Figure 1.16 that displays the number of years of attendance per cohort). We can conclude that the most likely assumption is that there is a time gap between the SHS graduation of the students and the year they start attending a post-secondary institution.

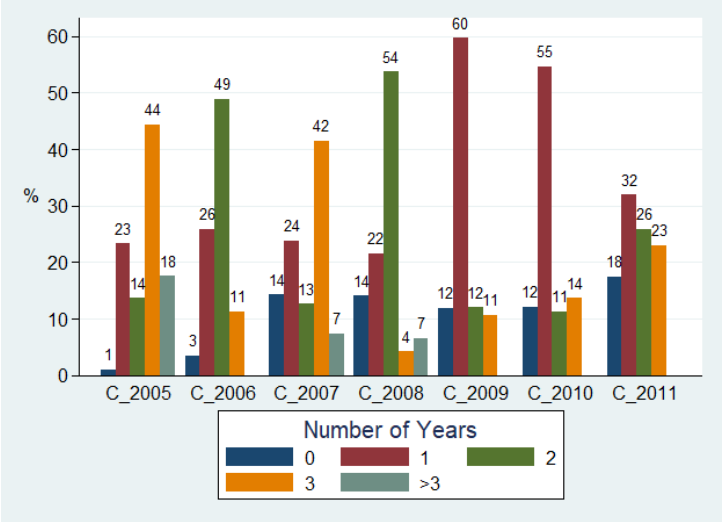
Figure 1.17 presents the distribution of the respondents according to the type of post-secondary institution they are currently attending. The ranking is consistent with the type of the institution applied. Around 41% of the students attend a university while 15% of them attend a polytechnic college, 15% are in teacher training and 15% are in nursing education. There are more male

FIGURE 1.14 Percentage of the the Respondents Who Are Currently Post-Secondary Students



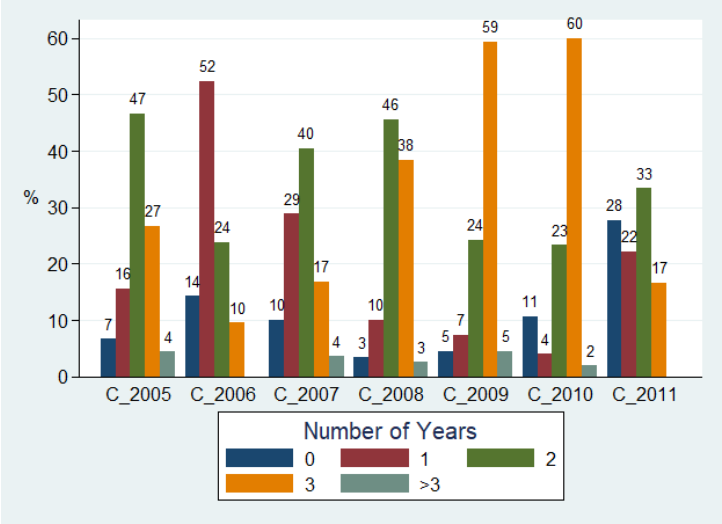
Note : Several observations per respondent (n = 14 885). 'In School' waves are excluded. The C_{2005} and C_{2006} cohorts were not interviewed in 2018. Anne Duplantier (2021)

FIGURE 1.15 Distribution of the Respondents According to the Time between SHS Graduation and Attendance of a Post-Secondary Institution



Note : Several observations per respondent (n = 6075). Anne Duplantier (2021)

FIGURE 1.16 Distribution of the Respondents According to the Time of Attendance of a Post-Secondary Institution

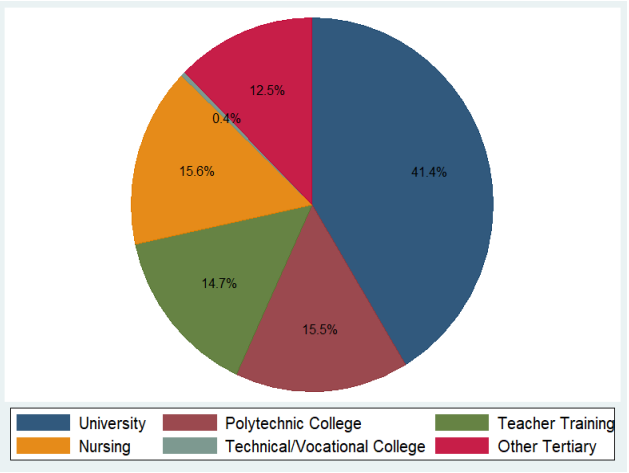


Note : Several observations per respondent (n = 1274). Anne Duplantier (2021)

students attending university (47.5%) than girls (32%). Moreover, over the 3836 respondents attending a post-secondary institution, 1086 boys attend university versus 501 girls - meaning boys are twice as likely to attend university as girls are. The proportion of students attending a teacher training institution is similar for boys and girls (15%) but there are slightly more boys attending a polytechnic college than girls. For the nursing schooling, the difference is much higher : 27.5% of girls are attending nursing schooling compared to 7.5% of boys. Gender differences are statistically significant and were quite expected. In general, there are more girls in fields such as nursing, and more boys choosing to attend scientific schools like polytechnic colleges.

Figure 1.18 presents the percentage of the respondents who have ever attended a post-secondary institution, by cohort and year - Graph (a) - and by gender for the 2018 year - Graph (b). Consistently, the percentage increases over time for each cohort (Graph (a)). In 2014, 78% and 73% of the oldest respondents from the C_{2005} and C_{2006} cohorts have ever attended a post-secondary institution. In 2018, 75% and 66% of the respondents from the C_{2007} and C_{2008} cohorts have ever attended a tertiary institution, and a little more than half of the respondents from the C_{2009} and C_{2010} cohorts. The youngest respondents from C_{2011} cohort have a probability of 45% to have ever attended a post-secondary institution while they graduated from SHS since 2014. In Graph (b), there are slight differences of attendance according to the gender. In each cohort except C_{2011}, girls have a slightly higher probability to have ever attended (in 2018) a post-secondary institution than boys but the differences are not statistically significant, except for the cohort C_{2008} where the difference is larger between girls and boys.

(a) Whole Population



(b) By Gender

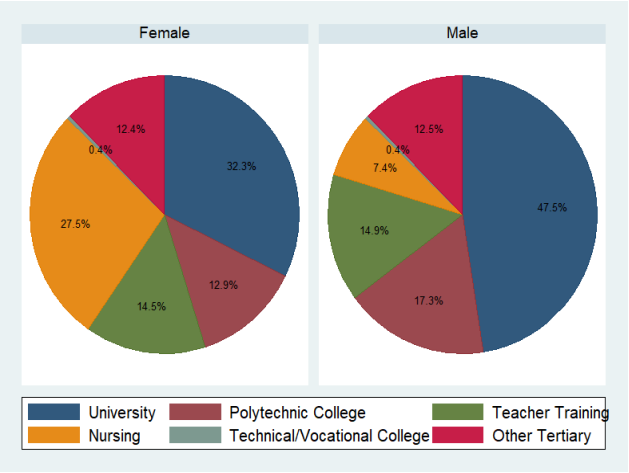


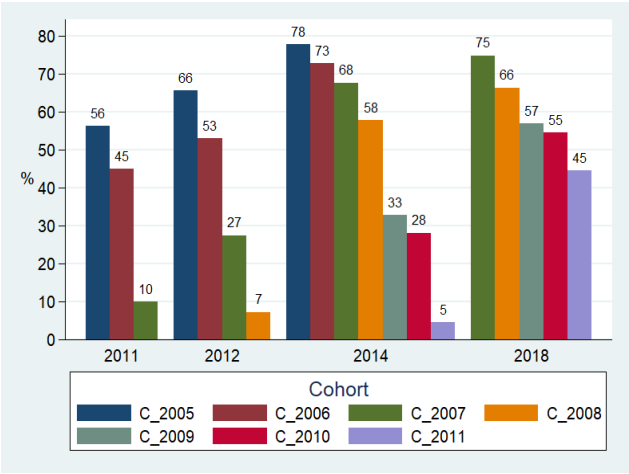
FIGURE 1.17 Distribution of the Respondents According to the Type of the Institution They are Currently Attending

Note : One observation per respondent (n = 3836). Gender difference's χ^2 test p-value = 0.000. Anne Duplantier (2021)

However, this is not representative of the trends in the whole population. The report from Ghana Statistical Service et al. (2015) finds that only 3.9% of the women and 7.5% of the men attained more than secondary education in 2014. Interestingly, looking at the whole population, women have less access to post-secondary education. But looking at the population of youth that went to secondary school, girls have a higher probability to have ever attended a post-secondary institution than boys. This suggests that once girls have access to secondary education, they become more likely to access post-secondary education.

Figures 1.19 and 1.20 show the percentage of the respondents who graduated and dropped-out from a post-secondary institution by gender. In the C_{2005} and C_{2007} cohorts, almost half of the respondents graduated from tertiary, while it is slightly less (39% and 36%) for the C_{2006} and C_{2008} cohorts. Regarding the C_{2009} and C_{2010} cohorts, only 17% and 13% of the respondents graduated from a post-secondary institution, though they both graduated from SHS since 2013. It is even higher for the C_{2011} cohort, among whom almost nobody graduated (1%). The students from the C_{2011} cohort are probably still attending a post-secondary institution, as they graduated in 2014 and there is a time gap between SHS graduation and entering into a post-secondary institution. Interestingly, girls have a higher probability to graduate than boys in every cohort, but the difference is not statistically significant for every cohort except C_{2005} and C_{2008}. The same trend appears for the probability to drop out : girls drop out more than boys in every cohort (but again the differences are not statistically significant). In general, around 10% of the students from the C_{2005}, C_{2006} and C_{2007} cohorts dropped out from a post-secondary institution. In the other cohorts, there are fewer dropped-out students (6% in the cohort C_{2008} and 3% for the C_{2009} and C_{2010} cohorts).

(a) All Years



(b) By Gender - 2018 Year

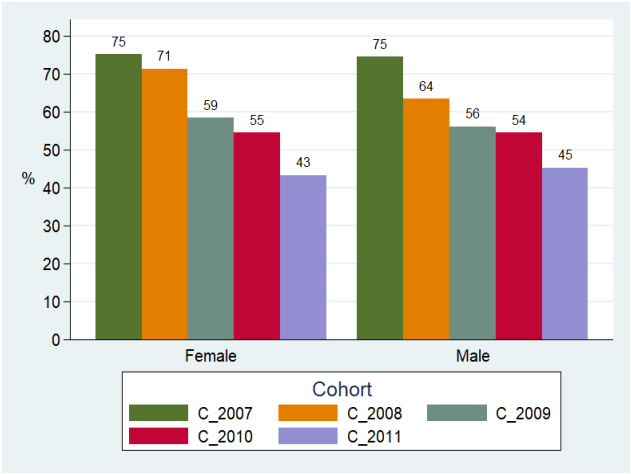
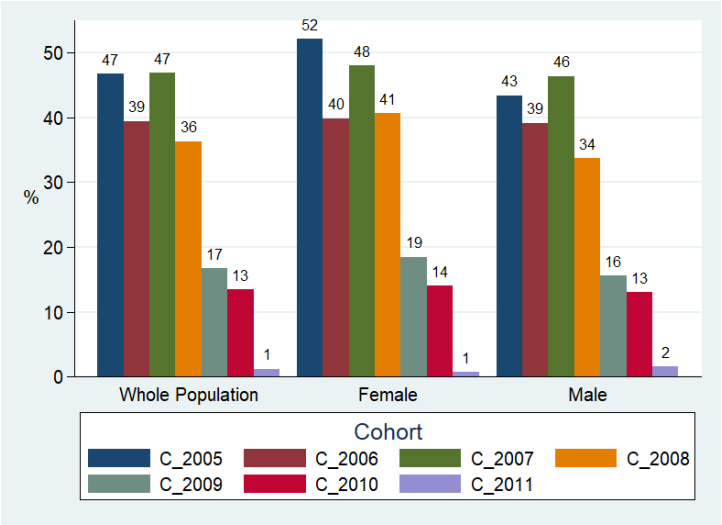


FIGURE 1.18 Percentage of the Respondents Who Ever Attended a Post-Secondary Institution

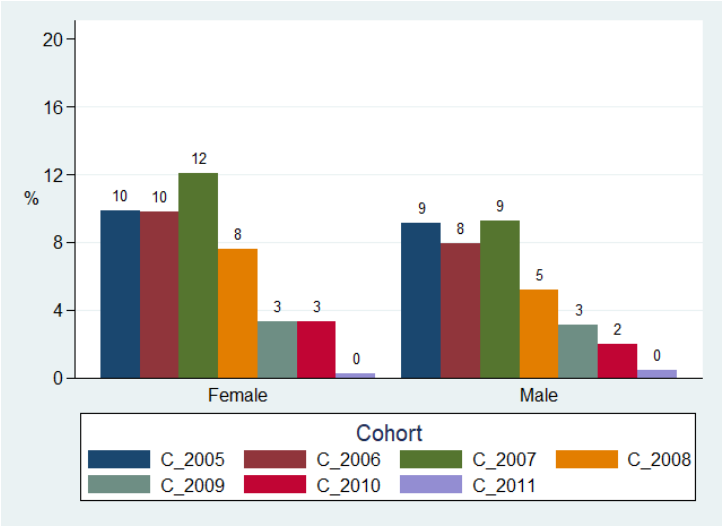
Note : The C_{2005} and C_{2006} cohorts were not interviewed in 2018. Several observations per respondent : in Graph (a) n = 14 908 and in Graph (b) n = 4260. Gender difference's χ^2 test p-value = 0.829 for C_{2007} ; 0.017 for C_{2008} ; 0.490 for C_{2009} ; 0.977 for C_{2010} ; 0.588 for C_{2011}. Anne Duplantier (2021)

FIGURE 1.19 Percentage of the Respondents Who Ever Graduated from Tertiary



Note : Percentage amongst all the respondents of the survey. One observation per respondent (n = 6568). Gender difference's χ^2 test p-value = 0.027 for C_{2005} ; 0.849 for C_{2006} ; 0.630 for C_{2007} ; 0.024 for C_{2008} ; 0.216 for C_{2009} ; 0.648 for C_{2010} ; 0.208 for C_{2011}. Anne Duplantier (2021)

FIGURE 1.20 Percentage of the Respondents Who Ever Dropped Out from Tertiary



Note : Percentage amongst all the respondents of the survey. One observation per respondent (n = 6568). Gender difference's χ^2 test p-value = 0.741 for C_{2005} ; 0.363 for C_{2006} ; 0.158 for C_{2007} ; 0.116 for C_{2008} ; 0.876 for C_{2009} ; 0.176 for C_{2010} ; 0.538 for C_{2011}. Anne Duplantier (2021)

1.4.5 Labor Market Characteristics

In the last part of this section, some outcomes of labor market are presented. Table 1.8 shows the distribution of occupation [19] of the respondents' parents according to their education level. The majority of the fathers without any type of education or with a primary education are farmers (75% and 53% respectively). Fathers with a JHS or SHS level of education have more diverse types of occupation : farming, government employment, self-employment (mainly in services) or other waged employment (between 36% and 12% for each type). Fathers with a post-secondary level of education are in majority government employees (38.6%) and teachers (26.7%). The situation is quite different for the mothers. For each type of education, the first occupation is always self-employed in trading (between 51% and 70%) except for mothers with a post-secondary education, who are government employees (30%) and teachers (30%). Interestingly, the proportion of farmers amongst the mothers decreases with the level of education (42% without education, 27.6% with primary level, 17% with JHS and 4.6% with SHS). These first statistics confirm that higher education in Ghana gives access to stabler and less vulnerable jobs, such as government employment or teaching, and makes it less likely to enter employment in agriculture.

The next tables and figures present the situation of the respondents in the labor market. Figure 1.21 shows the distribution of the respondents according to their primary activity in 2018. Amongst the oldest respondents (C_{2007} and C_{2008} cohorts), around 65% are working, between 15% and 19% are

19. The survey does not allow to measure informality. Thus, both informal and formal jobs are considered in this thesis, without distinction between the type of sectors.

TABLEAU 1.8 Occupation of the Parents According to Their Education

Level of Education	None	Primary	JHS	SHS	Post-Secondary
Father's Occupation for Majority of Life ($n = 2825$)					
Farmer (%)	75.10	53.13	36.34	20.46	3.07
Gov Employee (%)	3.26	7.29	13.77	16.75	38.65
Teacher (%)	0.19	0.0	1.82	4.75	26.69
Self-Empl (Services) (%)	7.47	16.67	19.94	17.46	8.13
Self-Empl (Trading) (%)	9.00	9.38	12.94	12.87	3.07
Self-Empl (Manufacturing) (%)	1.92	3.13	3.24	5.29	1.84
Other Wage Empl (%)	2.49	9.38	12.25	22.05	17.33
Unemployed (%)	0.38	0.0	0.10	0.18	0.15
Retired (%)	0.19	1.04	0.10	0.35	1.07
Number of Observations	522	96	988	567	652
Mother's Occupation for Majority of Life ($n = 2988$)					
Farmer (%)	41.80	27.64	16.99	4.60	1.87
Gov Employee (%)	0.34	0.41	1.78	7.51	30.22
Teacher (%)	0.11	0.0	1.01	4.12	29.85
Self-Empl (Services) (%)	1.14	2.85	5.07	9.69	5.22
Self-Empl (Trading) (%)	50.80	62.20	69.57	66.59	18.28
Self-Empl (Manufacturing) (%)	1.59	5.28	2.20	2.42	1.12
Other Wage Empl (%)	0.34	0.81	1.86	4.12	11.94
Unemployed (%)	3.87	0.81	1.52	0.97	1.49
Retired (%)	0.0	0.0	0.0	0.0	0.0
Number of Observations	878	246	1183	413	268

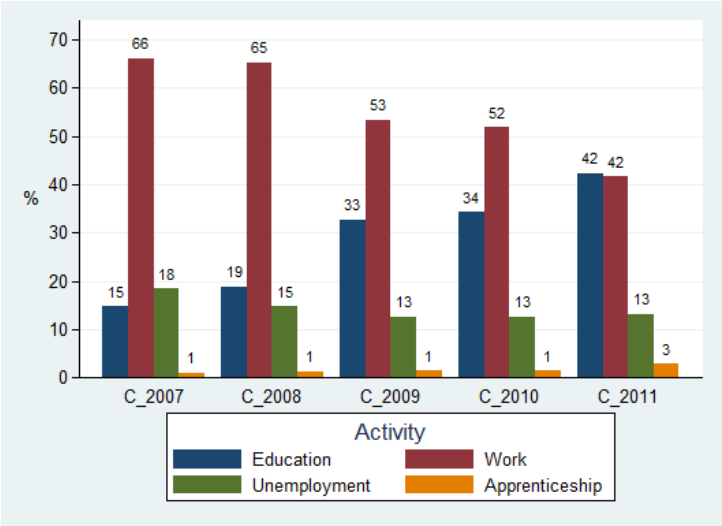
Note : Question about the occupation of the parents was asked in the 2011 and 2012 'Go Transition' waves. The percentages are computed in columns. For example, 75.10% of the fathers with no education are farmers. Anne Duplantier (2021)

studying at the post-secondary level and between 15% and 18% are unemployed. The proportion of unemployed amongst the youngest respondents (C_{2009} to C_{2011} cohorts) is close (13%) to the proportion amongst the oldest respondents, but there are more post-secondary students and fewer workers amongst the young respondents. The unemployment rates are quite consistent with the numbers from the World Development Indicators of the World Bank. In 2018, the unemployment rate in the total labor force aged 15-24 was 13.7%.

Figure 1.22 reports the descriptive statistics by gender. In every cohort, girls have a higher probability to be unemployed and a lower probability to work than boys (statistically significant differences for cohorts C_{2007}, C_{2008}, C_{2009} and C_{2010}), whereas the probability to study is quite the same (no statistically significant difference for every cohort). While there are more girls who graduated from post-secondary education, not many of them are actually working. These evidences seem to show that it might be less easy for educated girls to find employment compared to the educated boys. This is consistent with the finding that girls have a higher probability to apply to shorter programs in post-secondary institutions.

Table 1.9 reports the number of jobs the respondents had at the same time during the seven days before the interview and the number of worked hours by week conditional on working. The question was asked in the following way : "Have you done any work for pay, profit, or family gain or did you produce anything for barter or home use during the last seven days, even if it was for only one

FIGURE 1.21 Distribution of the Respondents According to the Primary Activity in 2018



Note : One observation per respondent (n = 5482).

(a) Female

(b) Male

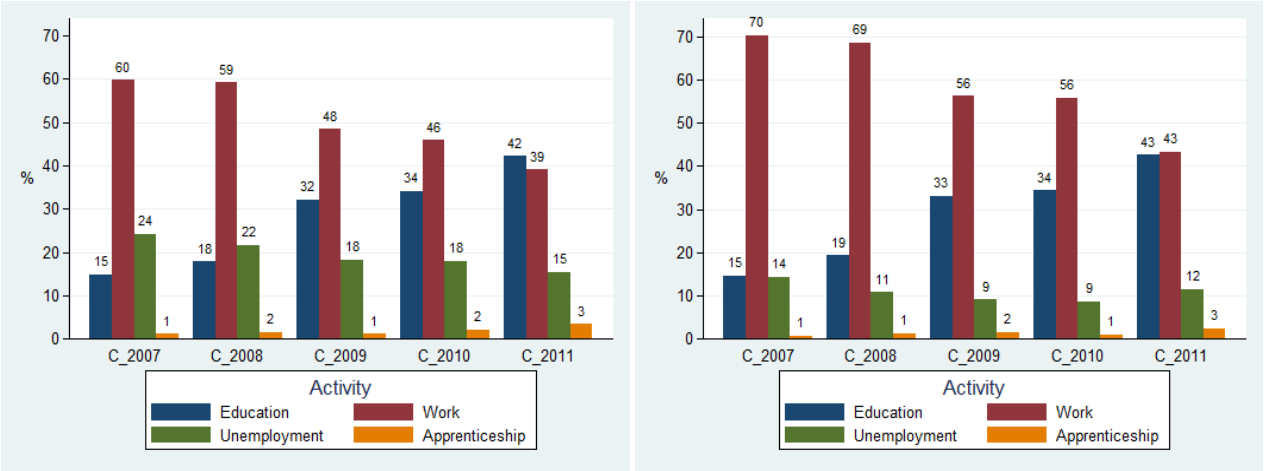


FIGURE 1.22 Distribution of the Respondents According to the Primary Activity in 2018

Note : One observation per respondent (n = 5482). Gender differences are statistically significant (at 5%) for the cohorts C_{2007}, C_{2008}, C_{2009} and C_{2010}, for the categories 'Work' and 'Unemployment'. Anne Duplantier (2021)

hour ?". On average, respondents had only one job and worked around 45 hours a week. There is no statistically significant difference between girls and boys concerning the number of worked hours a week. However, the number of jobs is statistically significantly lower for girls compared to boys, even if the difference is very small.

TABLEAU 1.9 Employment Characteristics

	(1) Whole	(2) Girls	(3) Boys	(4) P-value
Number of Jobs [$n = 3932$]	1.04	1.02	1.05	0.000
	(0.21)	(0.16)	(0.24)	
Number of Worked Hours [$n = 8440$]	45.98	45.86	46.02	0.352
	(19.30)	(19.22)	(19.34)	

Note : Several observations per respondent. Standard deviations are in parentheses. The question about the number of jobs was not asked in the 2014 'Go Transition' wave. The number of worked hours is conditional on working. Column (4) is the p-value of the t-test difference between Columns (2) and (3). Anne Duplantier (2021)

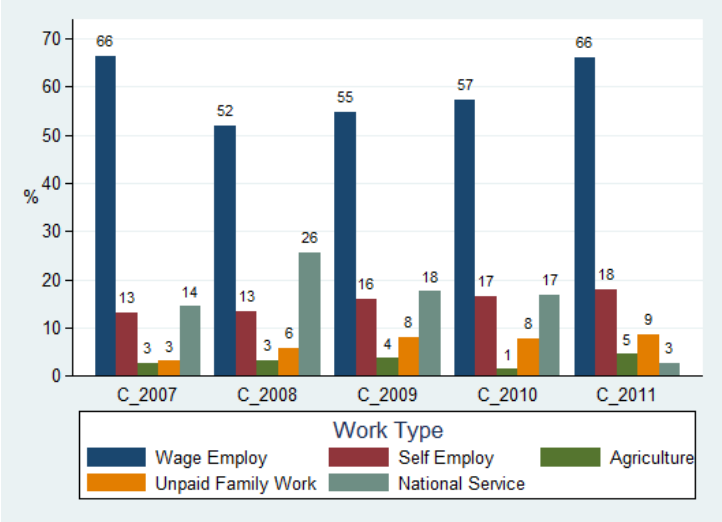
Figure 1.23 presents the different types of work (wage-employed, self-employed, agriculture, unpaid family work and national service [20]) held by the respondents who were working in 2018. Consistent with Table 1.8, the main type of occupation is waged employee (varying between 52% and 68% according to the cohorts). There are more self-employed amongst the youngest respondents (between 16 and 18% for the C_{2009} to C_{2011} cohorts). The proportion of unpaid family workers and workers in agriculture is very small in each cohort (between 1 and 9%). Interestingly, we found very different trends in types of work from the statistics from the World Development Indicators of the World Bank. In 2018, wage and salaried workers were 25.5% of the total employment versus 74.5% of self-employed. The trend is quite inverse for our educated population, suggesting that education gives a higher probability to access wage employed.

Figure 1.24 display the distribution of real weekly earnings. The distribution is very left-sided. This means that a high proportion of the population has low earnings and very few persons have high earnings. Moreover, the dispersion of earnings is quite high as most of the population earn less than 100 GHS per week whereas some respondents earn until 600 GHS per week.

Figure 1.25 shows the differences in real weekly wages between the regions for the year 2018. Over the whole sample of 7616 individuals, 25% (i.e. 1940 respondents) worked and earned an income in 2018. In average, the real wage was 88 GHS per week. Looking by regions, there are some differences (but not statistically significant). Real weekly wages are the lowest in Northern (80 GHS), a region where agriculture is dominant. Moreover, real weekly wages are the highest in Upper West, where the respondents earned on average 25% more than those in Northern (101 GHS

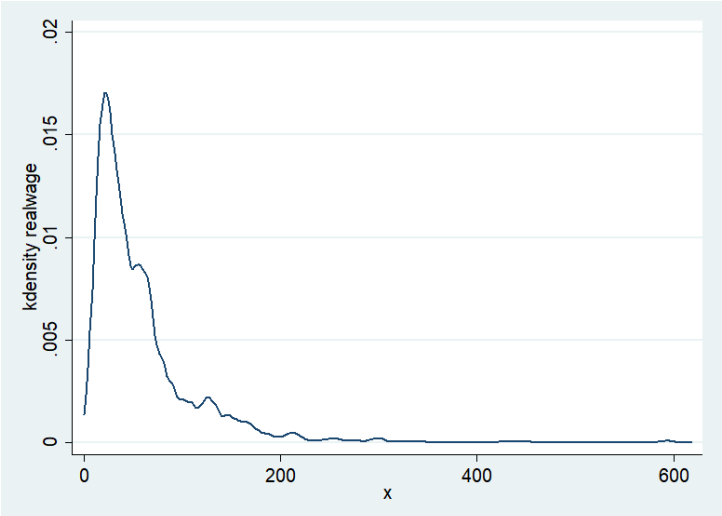
20. In Ghana, all Ghanaians graduating from public tertiary education (as a non-fee-paying) are required to serve the government by being posted during one year in public sectors that need human resources.

FIGURE 1.23 Distribution of the Respondents According to the Work Type in 2018



Note : One observation per respondent (n = 3104). Anne Duplantier (2021)

FIGURE 1.24 Distribution of real weekly wage



Note : Several observations per respondent (n = 7396). Anne Duplantier (2021)

per week). However, we must interpret this data carefully as Upper West has the smallest number of observations for wages (39 over 1940). The average wage in Greater Accra is around the national average. This can be explained by the fact that 34% of the observations we have on wages (657 over 1940) are located in Greater Accra.

Figures 1.26, 1.27 and 1.28 present the evolution of weekly real wages for different groups of the respondents over time (from 2008 to 2018). In these figures, information is recall data for the years 2008, 2009, 2010 and 2013.[21] This means that we asked the respondent in year y to remember information for previous years $y - n$. The wages have been deflated to allow comparison and take into account the inflation. Base year is 2011 and deflator data are from the World Development Indicators of the World Bank.

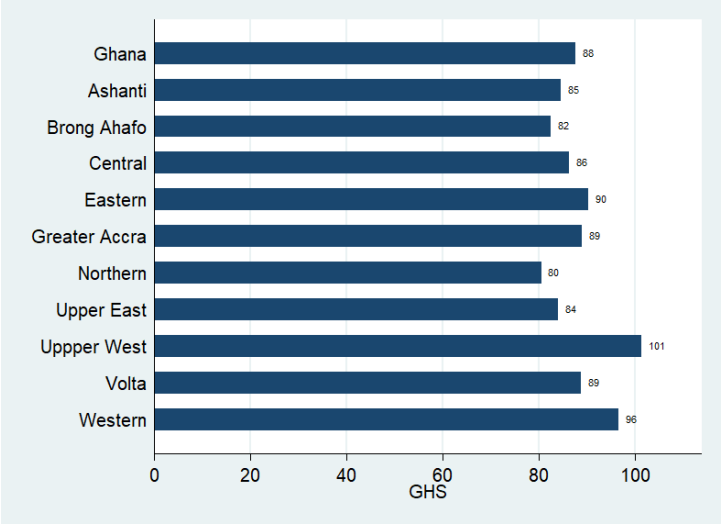
In Figure 1.26 (a), the average weekly real wages are presented by cohort. In general, wages are growing over time, likely because of experience and/or the level of education. The ranking of wages is quite consistent. For each cohort, wages increase over time. The C_{2007} cohort starts with an average weekly wage of 30 GHS and the average wage has more than tripled seven years after (105 GHS in 2018). For each year, the oldest students have higher wages and the average wage difference is quite important between the cohorts (the difference is statistically significant for each year except 2009). For example in 2014, the respondents from C_{2005} cohort have an average weekly wage of 88 GHS while the youngest respondents from C_{2011} cohort earn 35 GHS a week. Figure 1.26 (b) shows the evolution of the average weekly real wages according to the gender of the respondents. Whatever the year is, the male respondents earn on average higher wages than the female respondents (the difference is statistically significant for every year, except 2009 and 2011). The difference is generally the same over time and about 7 GHS a week (90 GHS versus 83 GHS), except in 2009 and 2010 where the gap is bigger than the other years.

Figure 1.27 depicts the evolution of the weekly real wages according to the post-secondary education attendance and the type of post-secondary institution attended. Figure 1.27 (a) compares the average wage between respondents who attended a post-secondary institution (excluding the current students[22]). During the first years, the average wage of the respondents without a post-secondary education is higher. This can be explained by the fact that in 2008 and 2009, many respondents are still attending or did not yet start attending a post-secondary institution at the time of the survey. It takes time to graduate and find a relevant employment correspondent to the education level. It is likely that in 2008 and 2009, the respondents who attended a post-secondary institution and started working did not graduate from a post-secondary institution and only attended one or two years before dropping out. However, starting in 2010, respondents with a post-secondary

21. 30 observations in 2011 and 39 in 2012 are also from recall data.

22. These respondents are excluded to better capture the effect of the post-secondary education, as it is less likely that a post-secondary student held full-time employment relevant to their education level while still studying.

FIGURE 1.25 Average Weekly Real Wage by Region in 2018



Note : One observation per respondent (n = 1940). The one-way 'analyse of variance' test (ANOVA test) is run to check if there are differences in mean wage between regions. The F statistic is 0.56 and the p-value is 0.833. Anne Duplantier (2021)

(a) By Cohort

(b) By Gender

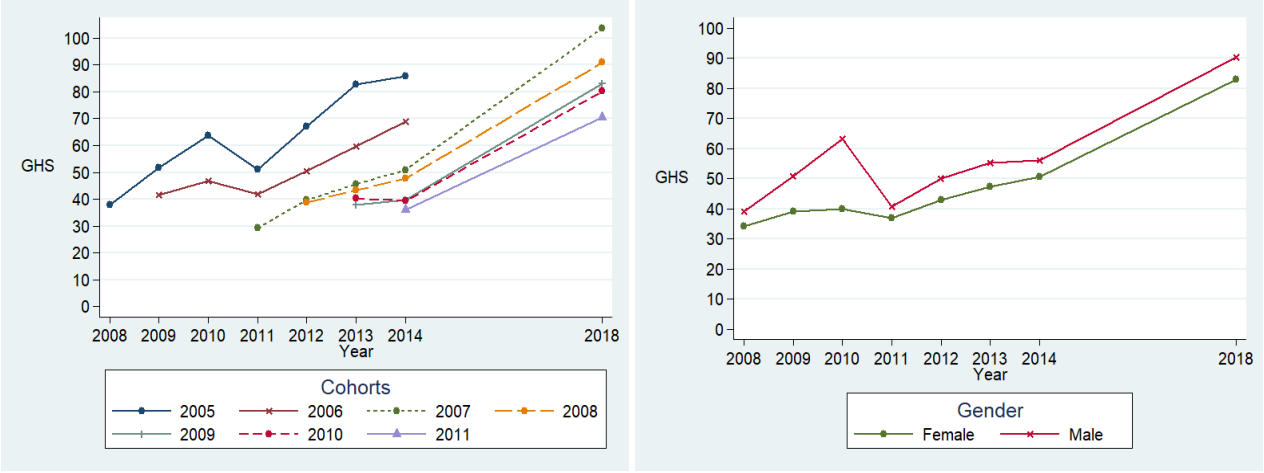


FIGURE 1.26 Evolution of the Average Weekly Real Wage

Note : In Figure (a) : There are several observations per respondent (n = 7396). The one-way ANOVA test is run to check if there are differences in mean wage between cohorts for each year. For 2009, the F statistic is 1.67 and the p-value is 0.196. For the other years, the p-value of the F statistic is inferior to 0.05. In Figure (b) : There are several observations per respondent (n = 7391). A t-test is run to check if there are gender differences in mean wage for each year. The difference is statistically significant for every year except 2008, 2009 and 2011. Anne Duplantier (2021)

education level have, on average, higher wages than those without a post-secondary education. In 2018, respondents who attended tertiary education earned 20 GHS more a week, which represents around 23% more than the respondents having never attended tertiary.

Figure 1.27 (b) presents the weekly deflated wages according to the type of post-secondary institution the respondent attended. Interestingly, respondents with nursing schooling or teacher training have the highest average weekly wage. In 2018, the average wage for a teacher was 120 GHS a week and 108 GHS for a nurse, while the respondents from university earned on average 88.5 GHS and from polytechnic college 85 GHS. Even for respondents who attended tertiary education, the teachers earn around 41% more than the respondents who attended polytechnic college.

Figure 1.28 (a) displays the average real wage according to the type of employment of the respondents. There are some variations according to the years, but in 2018, the self-employed respondents had the highest average weekly wage (120 GHS), followed by the waged employed (90 GHS), the respondents in national service (70 GHS) and the farmers (40 GHS). The self-employed respondents earn in average 200 percent more than the workers in agriculture. The differences are statistically significant every year. In order to compare with the number of worked hours, Figure 1.28 (b) presents the number of worked hours a week according to the type of employment of the respondents. The differences are statistically significant between the groups for every year except in 2009 and 2010. The self-employed respondents work on average 50 hours a week while the farmers work 35 hours. The ranking of the worked hours is quite consistent with the ranking of the wages. However, the differences in wages are much higher than the differences in terms of worked hours.

Figure 1.29 display the distribution of the worked hours per week. The distribution is more centered than the earnings distribution. The majority of the population work around 40 hours per week and a few persons work less than 25 hours per week. Moreover, the dispersion of earnings is quite high as there are still a lot of persons who work more than 50 hours per week.

Finally, Figure 1.30 presents the hourly real wage for the whole population and its evolution over time. In ten years, the average hourly real wage doubled. In 2008, the average hourly real wage equals 0.88 GHS whereas it is 1.93 GHS in 2018. However, on average, the hourly wage is very low. We computed the hourly wage by gender, by post-secondary institution attendance, by type of post-secondary institution and by type of work (as for previous figures about weekly wage). The trends between groups are the same than for the weekly wage.

(a) By Tertiary Education Attendance

(b) By Tertiary Institution Type

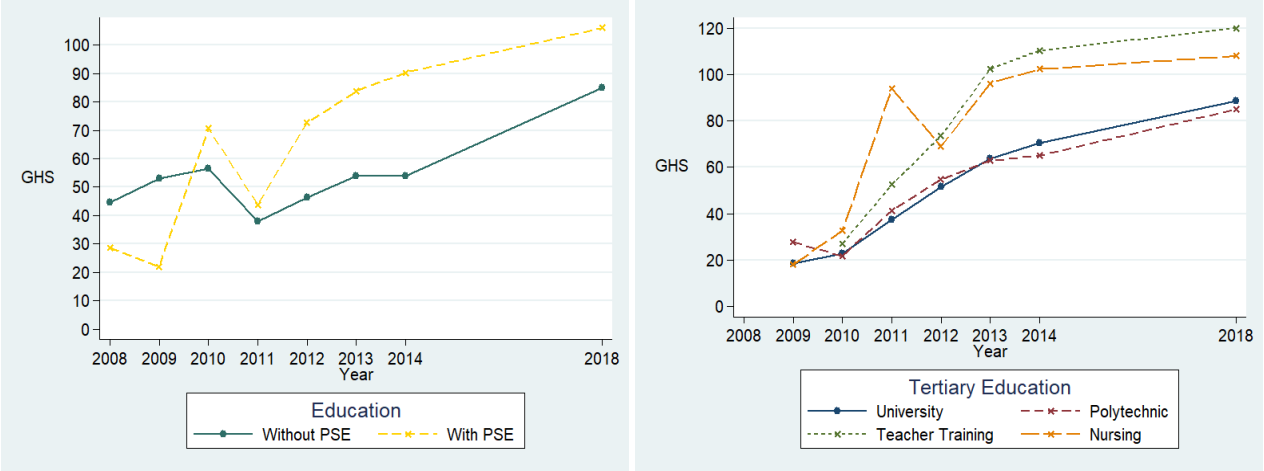


FIGURE 1.27 Evolution of the Average Weekly Real Wage by Education

Note : In Figure (a) : There are several observations per respondent (n = 6538). The current students are excluded. A t-test is run to check if there are group differences in mean wage for each year. The difference is statistically significant for 2012, 2013, 2014 and 2018 (and not for the other years). In Figure (b) : There are several observations per respondent (n = 1680). The one-way ANOVA test is run to check if there are group differences in mean wage between for each year. The differences are statistically significant for years 2009, 2010, 2014 and 2018. Anne Duplantier (2021)

(a) Average Weekly Real Wage

(b) Average Number of Hours

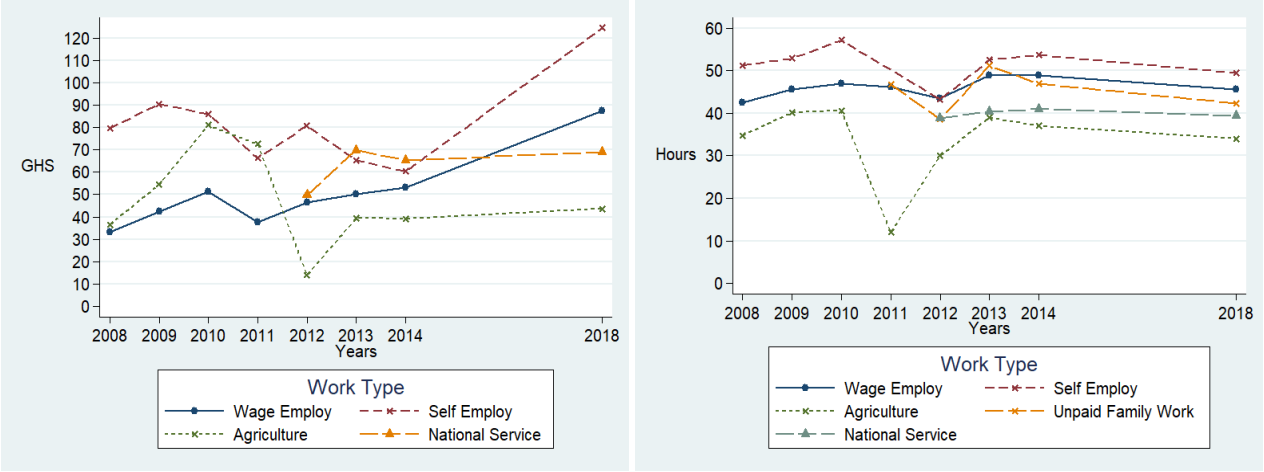
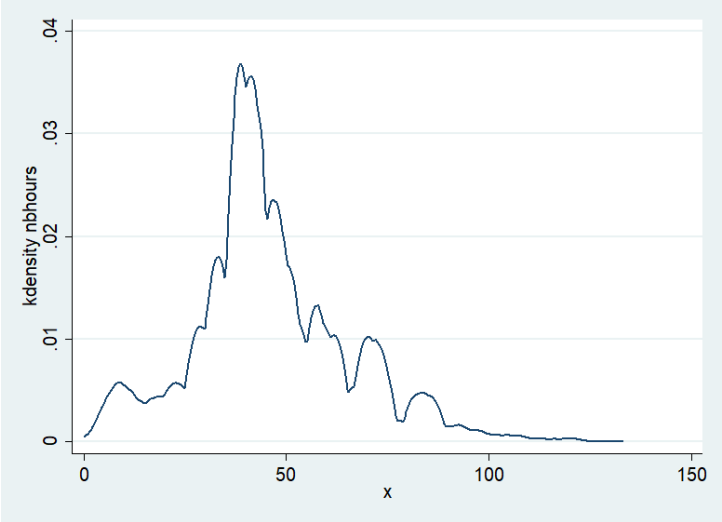


FIGURE 1.28 Average Weekly Real Wage and Average Number of Hours by Work Type

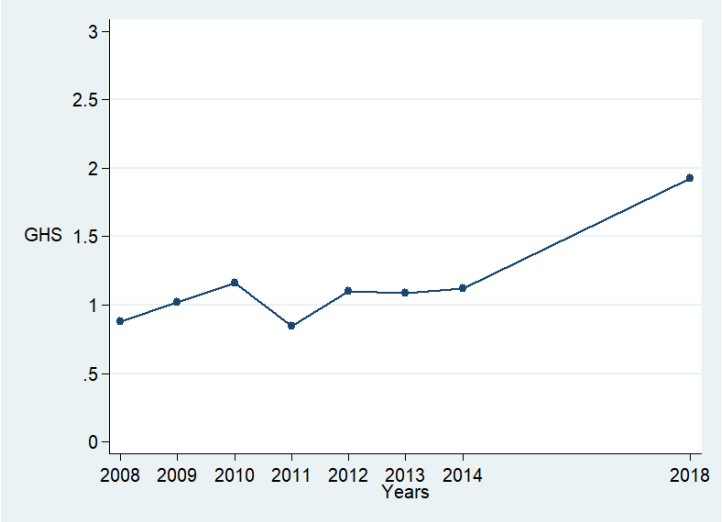
Note : In Figure (a) : There are several observations per respondent (n = 7106). The one-way ANOVA test is run to check if there are group differences in mean wage between for each year. The differences are statistically significant for all the years. In Figure (b) : There are several observations per respondent (n = 8105). The one-way ANOVA test is run to check if there are group differences in mean wage between for each year. The differences are statistically significant for all the years except 2009 and 2010. Anne Duplantier (2021)

FIGURE 1.29 Distribution of worked hours per week



Note : Several observations per respondent (n = 8440). Anne Duplantier (2021)

FIGURE 1.30 Average Hourly Real Wage



Note : Several observations per respondent (n = 5916). Anne Duplantier (2021)

1.5 CONCLUSION

This chapter presents the database I contributed to during my PhD years and use in my research. I first explained and discussed the survey methodologies, then depicted the situation of Ghanian youth in the educational system and the labor market.

The descriptive statistics can be summarized as follows. On average girls have less access to SHS than boys, but after SHS, girls have a slightly higher probability to attend a post-secondary institution. Moreover, girls attend more nursing schooling than boys, who attend more university than girls. Looking at the individual characteristics, it appears that girls come from less poor households, with relatively better living conditions than boys. It is consistent with the ranking of the reasons to not accept an offer : boys have a higher probability to not accept an offer after a post-secondary application for financial reasons than girls have.

With regards to the education system, access to SHS is costly, as the average fees are around 500 GHS, representing 21% of the per capita GDP in 2011. The average aggregate score of the terminal exam of secondary education (WASSCE) is around 18 (the first time the respondents pass it), which is slightly higher than the general minimum score required to enter into university. Interestingly, girls have a higher probability to graduate from tertiary but also to drop out of post-secondary education.

Finally, on the labor market outcomes, I find that girls have a higher probability to be unemployed than boys. While there are more girls who graduated from post-secondary educations, fewer of these enter work. It might be less easy for educated girls to find an employment compared to the educated boys. There is a premium to post-secondary education, as the respondents who attended a tertiary institution earn on average more than others.

Graduates of nursing schooling and teacher training, which are more applied and less academic education than university and polytechnic college, get the higher average wages. Chapter 3 will examine whether respondents' expectations on earnings match with this observation. I also find that there are differences in the weekly incomes according to the region of residence. We will study these regional differences of wages more deeply in Chapter 2. We will explore whether these differences are a driver of the Ghana's internal migration.

CHAPITRE 2
THE INTERNAL MIGRATION DECISIONS OF GHANAIAN YOUTH

Résumé

Si la migration internationale a suscité beaucoup d'attention au cours des dernières années, la migration interne est moins au centre de l'intérêt. Pourtant, ce type de migration est le plus courant dans les pays en développement. Pour de nombreux migrants en Afrique subsaharienne, la migration internationale n'est pas possible, principalement en raison des coûts et des risques élevés qui y sont associés. C'est pourquoi les individus, y compris les jeunes instruits, migrent souvent dans leur pays d'origine pour des raisons économiques et professionnelles. Ce chapitre se concentre sur la migration interne des jeunes éduqués au Ghana et analyse la relation entre les décisions de migration et les écarts de salaires entre les régions ghanéennes. La principale hypothèse sous-jacente est que les migrant fondent, au moins en partie, leur décision de migrer ainsi que le choix de destination sur leur probabilité d'emploi et leur revenu espéré à destination. Une base de données de source primaire et un modèle de choix sont utilisés pour examiner comment la probabilité d'emploi et les revenus attendus affectent les décisions de migration. Les résultats indiquent que les revenus espérés au niveau régional sont significativement liés à la probabilité de migrer. En outre, la variabilité des revenus est également considérée comme un facteur déterminant de la décision de migration, ce qui suggère que les migrants internes au Ghana ne sont pas réticents au risque. Enfin, les jeunes sortant de l'école secondaire qui sont nés dans une région rurale ont une plus faible probabilité de migrer, alors que la probabilité de migrer est positivement liée à l'éducation de la mère et aux capacités individuelles du jeune.

2.1 INTRODUCTION

According to De Vreyer et al. (2009), internal migration in West African countries is not a recent phenomenon and is widespread. Internal migration is generally less expensive than international migration. According to the United Nations Development Program, there were in 2009 approximately 740 million people who remained in their home country but had moved from their region of birth - about 11% of the world population. In Africa, more than 113.5 million individuals - 12.5% of the population - are considered as internal migrants (United Nations Development Program, 2009).

Internal migration is often the option chosen by educated individuals who are looking for employment opportunities. In Ghana, specifically, educated youth are particularly mobile. According to our survey, 31.5% of senior high school (SHS) students [1] lived outside of their region of birth in 2011/12. This chapter investigates the internal migration decisions of former SHS students. We examine how regional disparities within Ghana affect migration decisions of educated youth. We focus on the role that both mean expected income and income variability play in migration decisions and whether risk aversion and expectations can help explain migration patterns.

We answer these questions by using primary survey data collected in 2011/12 from students who graduated from, or dropped out [2] of, 136 [3] randomly selected public SHSs in Ghana between 2008 and 2012. The survey provides unique information on respondents' education outcomes and career expectations. We define internal migration as a change in the region of residence within Ghana. We use a behavioral choice model estimated with a mixed logit model [4] to investigate internal migration decisions.

This chapter contributes to the literature on internal migration in West African countries (Awumbila et al., 2008) by viewing migration as an individual decision and considering the income differences between regions as a determinant of internal migration. Regarding the methodology, we use a choice model estimated with a mixed logit overcoming some of the limitations of the standard logit model, such as the assumption of independence of irrelevant alternatives. Moreover, we contribute to the literature by focusing on educated youth, a population that is not frequently studied yet is crucial for the development of a country (as discussed in the Chapter 1). Finally, we integrate risk in the model of migration decisions by studying whether differences in variability in

1. As explained in Chapter 1, the Ghanaian education system consists of six years of primary school followed by three years of junior high school (JHS) which is the last stage of free and universal basic education for our sample. At the end of JHS there is a terminal examination, the Basic Education Certificate Examination, which is used to place pupils in SHS. Access to SHS is limited, with a transition rate from JHS3 to SHS1 of 50% (The Ministry of Education, 2012).

2. In this chapter, we use the term 'graduates' to refer to both SHS graduates and the 2.24% of the sample that did not successfully complete SHS.

3. Of the 515 public SHSs in Ghana in 2011.

4. Developed by McFadden and Train (2000).

incomes between regions attract migrants.

The findings can be summarized as follows. A higher average income in neighboring regions increases the migration of youth with an SHS education level. After controlling for regional mean income, income variability also stimulates migration, suggesting that migrants are risk-loving or have a high assessment of their success in the labor market. Finally, we find that migration is less common among youth born in rural areas and that it is positively correlated with the mother's level of education and a measure of individual ability.

The remainder of the chapter is structured as follows. Section 2 reviews the internal migration literature. Section 3 presents the data and descriptive statistics, while Section 4 discusses the methodology employed. Results are reported and discussed in Section 5, and Section 6 presents the concluding remarks.

2.2 LITERATURE REVIEW

2.2.1 Internal Migration

The economic literature has focused extensively on international migration, [5] but evidences on internal migration are spares - despite this being a widespread phenomenon. Leaving one's country is a very costly investment - often with legal restrictions. International migration is not a feasible option for most, and internal migration appears much more accessible. Indeed, Lewis (1954) highlights the importance of rural-urban migration in the development process. Deshingkar and Grimm (2005) emphasize that internal migration is increasingly common in developing countries. According to the latter authors, internal migration plays an important role in poverty alleviation and contributes to economic development, as it is often a key survival strategy for the poorest populations. However, migration is also desired by, and accessible to, a less poor and more educated segment of society seeking a higher return on their investment in education.

There are numerous examples in the literature of evidence of the link between education and migration. For instance, Miguel and Hamory (2009) and Du et al. (2005) find that higher education and skill levels increase the probability of migration in Kenya and China, respectively. On the contrary, Hare (1999) finds that the number of years of schooling have no impact on the probability of migration in rural China. Most studies find that single young men with no dependents are most likely to migrate (Hare, 1999; Du et al., 2005; Gray, 2009), although Deshingkar and Grimm (2005) find a pronounced feminization of migration in recent years, especially in South America and Southeast Asia. Moreover, decreasing opportunities in farming, high population densities and

5. For example with the works of Adams and Page (2005); Bertoli and Marchetta (2014); McKenzie and Sasin (2007) and Mayda (2010).

an unequal distribution of agricultural land are positively correlated with out-migration from rural areas (Deshingkar and Grimm, 2005).

The decision to migrate, and if so, to where, is often complex and may be based on many different factors. Amongst them are economic considerations. Rational individuals compare the expected benefits, both financial and non-financial, with the costs of migration. Lall et al. (2009) and Zhang and Shunfeng (2003) find that wage differentials between rural and urban regions create an incentive for migrants to leave the countryside. This study is consistent with the traditional theory of Harris and Todaro (1970) and Lewis (1954). However, Fafchamps and Shilpi (2013) find that inter-regional mean wage differentials do not affect internal migration in Nepal, suggesting that differences in spending and consumption levels between regions play a role in migrants' destination choices. Fafchamps and Shilpi (2013) examine subjective satisfaction derived from the consumption of food, clothing, housing, healthcare and education. In all of these areas, the level of subjective satisfaction is higher in the destination region compared to the region of origin and other alternative destinations.

Subjective satisfaction captures some facts that simple wage differentials cannot, such as psychological effects. One can imagine a situation where migration leads to a higher wage but lower subjective welfare, as distance to home and familial roots, cultural differences and the need to learn a new language can negatively affect the welfare of migrants. The social network of migrants at the destination is another key factor in migration decisions (Mora and Taylor, 2006). Fafchamps and Shilpi (2013) find, for example, that Nepalese migrants tend to limit their migration to areas where they share both language and ethnicity. Finally, Dudwick (2011) shows that migrants in Nepal are prepared to accept lower wages in order to access better government services.

Evidence on internal and international migration in West Africa is sparse, partly because of lack of data (Awumbila et al., 2008). Anarfi et al. (2003) discuss the determinants of rural-urban migration within Ghana. The authors suggest that high population growth rates over the past 30 years have increased the labor supply and put pressure on arable land, thus encouraging migration from the countryside to cities. Anarfi et al. (2003) find that large differences in poverty levels between the south and the north of Ghana are increasing the prevalence of internal migration. The coastal zone is the most industrialized and urbanized, attracting the majority of internal migrants. The center of the country (the Ashanti region), is rich in agriculture, forestry and mining. These regions received the majority of the internal migrants from the north in the 1990s (Anarfi et al., 2003).

Tsegai (2007) also studies internal migration in Ghana but focuses on the Volta region. Following the New Economics of Labor Migration theory, [6] Tsegai (2007) assumes that the migration decision is a household decision. Using household survey data collected in 2001 from 500 households, he

6. Cf Stark and Bloom (1985).

studies the determinants of migration decisions. More precisely, Tsegai (2007) investigates the roles of income differences between migrant and non-migrant households. A probit model is employed and the selectivity bias is corrected using the Heckman procedure, whose one limit is to hold only if there is no selection on unobservables. Tsegai (2007) shows that migrant households have higher income than households without a migrant, and that these differences are a determinant of internal migration in the Volta basin of Ghana. Other factors like migration experience, household size, education and social capital are also found to explain migration decisions.

2.2.2 Risk and Migration

According to Stark (1981), risk attitudes play an important role in the migration behavior of both individuals and households. Todaro (1969) argues that rural incomes are without risk and assumes that workers are risk-neutral. Lucas (1997) questions these assumptions and suggests as a more realistic assumption that migrants are risk-averse. Indeed, expected benefits of migration are often uncertain (Stark, 1981), and the initial risks are high and costly (transportation, information and psychological costs) (Gibson and McKenzie, 2011). However, Bonin et al. (2009) note that economic theory has no clear prediction on attitudes of migrants toward risk, highlighting a lack of consensus on the relationship between migration and risk.

On the one hand, standard migration models, such as Heitmueller (2005) predict that the probability of migration is negatively related to risk aversion. Empirically, Chiswick (1978), Todaro (1980), and Constant and Zimmermann (2006) find that international migrants are more risk-loving than natives in both destination and origin countries. Using panel data including information about incomes, employment and risk attitude in Germany (2000-05), Jaeger et al. (2010) also find that individuals who are more risk-loving are more likely to migrate between labor markets. They measure risk aversion by asking respondents to evaluate their willingness to take risks on a scale of 11 categories [7] and define migration as a move from one region to another inside of Germany. Guiso and Paiella (2004) find that risk aversion predicts individual behavior in Italy, in particular occupational choice and migration decisions. They define a migrant as an individual who moved from their birth region to another region in Italy. They find no statistically significant difference between the probability of migration between risk-loving and risk-averse people. However, amongst risk-averse individuals, those who are more risk-averse have a lower probability of migrating.

There is little evidence regarding the link between migration and risk in developing countries. Gibson and McKenzie (2011) followed successful students in secondary school from three Pacific countries (Tonga, New Zealand and Papua New Guinea) from 1976 to 2004. They find that risk-

7. Zero indicates that the respondent is completely unwilling to take risks and ten indicates that they are completely willing to take risks.

seeking and more patient individuals are more likely to migrate. In China, Akgüç et al. (2016) find that risk tolerance is positively related to a higher probability of rural-urban migration. While the previous studies rely on self-reported measures of risk tolerance, Hao et al. (2016) and Goldbach and Schlüter (2018) use incentivized field experiments in order to elicit the true preferences of the respondents. Goldbach and Schlüter (2018) compare the preferences of internal migrants [8] and non-migrants from two localities in Indonesia and Ghana. They highlight that in addition to the standard factors of migration such as age, education or employment status, risk and time preferences are major determinants. In both regions, risk-averse individuals have a lower probability of migrating than their risk-loving counterparts. Goldbach and Schlüter (2018) conclude that migrants are, on average, younger, better educated and likely more risk-tolerant and patient. In China, Hao et al. (2016) find relatively mixed results : Chinese migrants are more competitive than non-migrants, but they do not have different risk preferences when strategic uncertainty [9] is absent, i.e. where there is no simultaneous competition with other players.

On the other hand, Katz and Stark (1986) reject the hypothesis that rural-to-urban migrants are risk-loving. In their theoretical work, they show that risk-averse individuals migrate because they assess the expected risk over their lifetime. The authors suggest that the variability of income from urban employment decreases over time for a migrant and may be lower than the risk of agricultural production. Therefore, migration is compatible with global risk aversion if individuals smooth their expected risk over their lifetime. Empirically, Bonin et al. (2009) find that, in Germany, first-generation migrants are more risk-averse than natives while the risk preferences of second generation migrants, i.e. born in Germany but with migrant parents, are equal to those of the native population. These findings suggest that, depending on preferences, migrants may be more or less risk-loving than natives, on average.

The risk literature is almost unanimous regarding the relationship between gender and risk. Most studies find that women are more risk-averse than men (Borghans et al., 2009; Hartog et al., 2002; Agnew et al., 2008; Ding et al., 2010; Eckel and Grossman, 2008; Barsky et al., 1997; Donkers et al., 1999). Only Harrison et al. (2007) find no gender difference in risk attitudes (in Denmark). However, in a large review of the empirical literature, Nelson (2015) further investigates the finding that women are more risk-averse than men. Amongst the empirical studies, 35 have found that women have a higher preference for risk, while many of the studies suggesting that women are more risk-averse than men lack statistical significance. In studies where a statistically significantly higher preference for risk amongst men is found, the size of the difference is generally less than

8. About 44% and 53% of the households who participated in the survey in Indonesia and Ghana, respectively, had at least one migrant. Moreover, only 4% and 7% of migrants from Indonesia and Ghana, respectively, actually left the country.

9. The strategic uncertainty is measured by “the willingness to compete for a limited number of prizes when others’ decisions are simultaneous". The authors use a market entry game to elicit that concept.

one standard deviation. Therefore, Nelson (2015) suggests that the conclusion of important gender differences in risk aversion in the existing literature is an overstatement.

As this chapter assesses migration decisions of educated youth, we briefly discuss the literature examining the relationship between risk and education. In their empirical work, Hartog et al. (2002) study the relationship between risk aversion and individual socio-economic and demographic characteristics. Using three different Dutch databases, the authors find that a rise in income increases the probability of an individual being either risk-neutral or risk-loving. The authors also find that education decreases risk aversion. These findings are consistent with the results of Schwartz (1976), Binswanger (1980), Binswanger (1981), and Donkers et al. (1999) which suggest that the more educated individuals are, the less risk-averse they are. However, Harrison et al. (2007) find that, in Denmark, individuals with vocational training and higher education have a higher probability of being risk-averse than people with lower education levels.

2.2.3 Position in the Literature

First, this chapter fills a gap in the literature and data on internal migration in West African countries (Awumbila et al., 2008). Indeed, the topic of internal migration in SSA is often difficult to address because of a lack of data. Bringing a rich dataset that allows to study the different steps of migration of each respondent is very valuable. Moreover, most of the studies on internal migration in SSA deal with rural-urban migration. In this chapter, we overcome this limit by focusing on migration between regions of Ghana without restricting to rural-urban moves. It allows to capture a larger picture of internal migration in the country by including in the analysis all the urban-urban migration.

Following Anarfi et al. (2003) and Tsegai (2007), we focus on internal migration in Ghana and consider income differences as a determinant of the migration decision. However, this chapter differs from Tsegai (2007) on several points. Firstly we provide a large database with a sample at the national level, while Tsegai (2007) focus on the Volta region. In this study, the sample size is 501 observations whereas our subset of data has more than 2300 observations.

Secondly, unlike Tsegai (2007), who considers migration as a household decision, we consider the latter as an individual decision. Instead of focusing on income differences between migrants and non-migrants, we consider the income differences between regions as a determinant of internal migration. Moreover, we use an external database to measure the average income by region. The GLSS data is a credible source of data, with a higher sample size by region and a higher representativeness. This allows to capture the effect of general equilibrium that might occur if education individuals migrate to higher incomes.

Thirdly, this chapter and his database contributes to the literature by focusing on the educated youth. This is a population which is not frequently studied while is crucial for the development of a country. Indeed, youth represent a substantial economic advantage in SSA. However, this population faces also a significant challenge with a very high of unemployment. In 2011, the youth unemployment rate in Ghana was 65% according to the World Bank. [10] One solution for the young people searching for opportunities might be to migrate internally. Thus, understanding the link between internal migration of educated youth and differences of income across regions is a crucial stake for the country.

Finally, the chapter has several methodological contributions. One of them is to integrate risk into our model of migration decisions. Indeed, in addition of looking at the differences in average income between regions, we use differences in income variability between regions. It allows to understand if the educated internal migrants in Ghana are attracted or not by high variability of incomes. A second methodological strength of this chapter is to use a behavioral choice model estimated with a mixed logit. This allows to overcome some of the limitations of the very used standard logit model such as the assumption of independence of irrelevant alternatives. Moreover, both push and pull factors of migration are considered in this thesis. Indeed, average income in both destination and origin regions are included in the ratio of incomes. Therefore, it is not only the attractive income at destination that pulls the migrant outside of her region but the relationship between the income at origin and the one at destination. This is quite an unusual way to consider the role of income in the migration decision.

2.3 DATA AND DESCRIPTIVE STATISTICS

In this section, we first present the primary source of data and how we capture migration. The main characteristics of the respondents are also discussed, followed by their pattern of migration. In the second part of the section, we describe the data from where the regional average income are extracted and how the averages have been computed. Then the regional average income and the unemployment rates are presented according to different sub-groups of population, and we look at correlations between average income and migration rate.

10. https ://web.archive.org/web/20120213132827/http ://www.ghanatoghana.com/Ghanahomepage/world-bank-assist-national-youth-employment-programme

2.3.1 SHS Graduates

The primary data come from the 'Go Transition' 2012 wave [11] of the Ghana Opportunities for Transitioning Senior High School Students survey (collected by the authors). The survey includes SHS graduates from 2008, 2009, 2011 and 2012. [12] This sample is representative of students that finished public SHS between 2008 and 2012. A large majority of respondents (79%) are aged between 17 and 23 years old. [13]

As this chapter focuses on economic migration, we restrict the sample to former students not attending university or a post-secondary institution at the time of the survey. In addition, individuals still attending SHS at the time of the survey are also removed from the analysis. Our final sample consists of 2311 individuals who were not in school at the time of the survey. [14]

We identify migration using the region of residence at three separate moments for each individual on the sample, which are (i) birth, (ii) SHS and (iii) time of interview (after SHS). A migrant is identified as an individual who was living at the time of the interview in a region that is neither their birth region nor a region in which they attended SHS. It corresponds to the last two cases of Table 2.1, which enumerates the different moves a respondent might have experienced. We try to capture the respondents who have migrated for economic reason and to not include all the migration decision which might come from studying or following their parents. Therefore, our definition of migrant does not consider individuals who attended SHS in a different region to where they were when interviewed but who returned to their birth region after SHS - Case 2 in Table 2.1 - as the motivation for this type of migration may be different from migration for economic reasons. Nor do we consider individuals who attended SHS in a different region to their birth region and stayed there - Case 3 in Table 2.1. Indeed, it is more likely that at this age and stage of life, the migration decision was made by their parents. [15]

Descriptive statistics are presented in Table 2.2. The average age at the time of the interview is 22 years old and 61.7% of respondents are male. On average, the time elapsed between SHS graduation and the time of the interview was 1.5 years ; the maximum being 4 years for the 2008 cohort and the minimum 4 months for the 2012 cohort. Less than 3% of the individuals are married. On average,

11. Precisely 2205 observations come from the 2012 round and 106 observations from the 2011 round (individuals who were not able to re-interview in 2012).

12. A reform of the education system in Ghana resulted in no graduation in 2010 and two cohorts graduating in 2011.

13. Around 87% of the respondents are aged from 19 to 24 years old.

14. Of that group, 97.11% had obtained their SHS diploma.

15. However, the analysis is not sensitive to the definition of migrant as when we include Case 3 in the definition of migration, the results are very similar.

FIGURE 2.1 Regions of Ghana

Source : emapsworld.com

TABLEAU 2.1 Definition of an Internal Migrant

Time	Birth	SHS	Now	Moves	Migrant
Case 1	$Region_A$	$Region_A$	$Region_A$	No move	No
Case 2	$Region_A$	$Region_B$	$Region_A$	Moved after birth and returned to birth region after SHS	No
Case 3	$Region_A$	$Region_B$	$Region_B$	Moved after birth and stayed in SHS region	No
Case 4	$Region_A$	$Region_A$	$Region_B$	No move after birth and moved after SHS	Yes
Case 5	$Region_A$	$Region_B$	$Region_C$	Moved after birth and moved after SHS	Yes

Anne Duplantier (2021)

TABLEAU 2.2 Differences in Characteristics Between Migrants and Non-Migrants

Variables	(1) Full sample	(2) Migrants	(3) Non-migrants	(4) Difference	(5) Standard error
Individual characteristics uninfluenced by migration decisions					
Male (%)	61.67	63.87	61.06	2.81	(2.45)
Age (average)	21.95	22.21	21.87	0.33***	(0.11)
	(2.23)	(2.33)	(2.19)		
Born in rural area (%)	59.57	64.63	58.16	6.46**	(2.54)
At least one parent with SHS (%)	35.78	34.46	36.15	-1.69	(2.42)
Father with no education (%)	19.63	22.55	18.83	3.72*	(2.14)
Mother with no education (%)	31.98	32.67	31.79	0.87	(2.47)
Cognitive ability	40.02	39.62	40.13	-0.51	(1.26)
(% questions correct)	(23.59)	(23.39)	(23.66)		
Average years since SHS	1.44	1.69	1.37	0.32***	(0.07)
	(1.47)	(1.54)	(1.44)		
Boarding during SHS (%)	52.10	51.79	52.18	-0.39	(2.52)
Number of siblings (average)	3.83	3.86	3.82	0.04	(0.17)
	(2.13)	(2.42)	(2.05)		
Current individual characteristics					
Married (%)	2.47	3.02	2.31	0.70	(0.78)
Worked last 7 days (%)	43.14	47.77	41.84	5.94**	(2.52)
Type of current work					
Wage employed (%)	91.64	94.86	90.66	χ^2 test p-value :	
Self employed (%)	3.47	2.80	3.68	0.112	
Agriculture (%)	4.89	2.34	5.66		
Months worked per year (average)					
Graduated SHS before 2012	7.80	7.43	7.92	-0.48	(0.35)
	(4.08)	(4.17)	(4.05)		
Graduated SHS in 2012	2.84	2.76	2.86	-0.11	(0.27)
	(1.78)	(1.47)	(1.87)		
Days worked per week (average)	5.46	5.58	5.43	0.15**	(0.06)
	(0.87)	(0.91)	(0.86)		
Hours worked per week (average)	41.21	46.56	39.63	6.93***	(1.60)
	(21.40)	(21.53)	(21.12)		
Average weekly wage in 2010 (GHS)	62.54	71.45	59.89	11.56	(14.79)
	(92.66)	(87.43)	(94.25)		
Searched for work last 2 months (%)	12.31	13.51	11.99	1.56	(2.29)
Apprenticeship (%)	2.29	1.99	2.38	-0.38	(0.75)
SHS characteristics					
Migration network per grade (%)	21.57	27.96	19.80	8.16***	(1.24)
	(24.73)	(27.21)	(23.71)		
Average distance SHS-Accra (km)	221.08	218.53	221.78	-3.25	(7.64)
	(151.54)	(157.37)	(149.92)		
Number of observations	2311	502	1809		
Wave of interview					
2011 (%)	4.59	4.78	4.53	0.25	(1.05)
2012 (%)	95.41	95.22	95.47	-0.25	(1.05)

Note : *** p<0.01, ** p<0.05, * p<0.1. Standard deviations are in parentheses. The weekly wage is conditioned by having worked in 2010. The status 'Searched for work last 2 months' is conditioned by having not worked last 7 days. For 'Searched for work last 2 months', the respondent was asked "Did you search for a job in July or August 2012 ?" for the 2012 survey which began in August, and "Did you search for a job in August or September 2011 ?" for the 2011 survey which began in September. Anne Duplantier (2021)

59.6% of the individuals were born in a rural area. [16]

As a proxy for migration networks, we construct the percentage of sampled students in each school (per grade) who have migrated, computed as a 'leave out one mean'. The underlying assumption is that a student takes into account the migration experience of students in their cohort of their SHS in the decision to migrate. On average, 21.5% of graduates have migrated in each grade in each school. [17] The schools are an average of 215 km from Accra ; the furthest school being 628 km from the capital.

We measure cognitive ability as the percentage of correct answers obtained on the 18 questions of the Klein test ; similar to the Raven's Progressive Matrices test [18] (used also in Borghans et al. (2009)). In our sample, individuals answered correctly to an average of 40% of questions. Approximately 36% of the respondents have at least one parent with at least an SHS level of education, while 32% of them have a mother with no formal education and 17% have a father with no formal education. The fathers of 18% of our participants have a post-secondary education while only 5% of mothers do.

TABLEAU 2.3 Differences in Location Between Migrants and Non-Migrants

	(1) Full	(2)	(3)	(4)	(5) Standard
Current region (%)	sample	Migrants	Non-migrants	Difference	error
Ashanti	19.50	16.14	20.45	-4.31**	(2.00)
Brong Ahafo	10.23	4.78	11.76	-6.98***	(1.52)
Central	7.70	6.97	7.90	-0.93	(1.35)
Eastern	6.30	1.79	7.56	-5.77***	(1.22)
Greater Accra	26.24	57.77	17.37	40.40***	(2.06)
Northern	7.08	2.59	8.35	-5.76***	(1.29)
Upper East	2.01	1.00	2.30	-1.30*	(0.71)
Upper West	2.54	1.39	2.86	-1.46*	(0.79)
Volta	10.67	1.20	13.33	-12.14***	(1.54)
Western	7.74	6.37	8.12	-1.75*	(1.35)
Observations	2311	502	1809		

Note : *** p<0.01, ** p<0.05, * p<0.1 Anne Duplantier (2021)

At the time of the survey, 22% of the respondents were living in a region that was neither their region of birth nor the region where they attended SHS. Columns (2)-(5) of Table 2.2 present the characteristics of migrants versus non-migrants. Migrants are on average slightly older (22.21 years

16. The survey does not provide a precise definition and uses a subjective measure : the respondent is asked whether their mother was living in a rural or urban area when they were born.

17. On average, students from each SHS have migrated in around two different regions (precisely 1.9 different regions).

18. This is a nonverbal test assessing the individual's ability to think, solve problems and learn.

old versus 21.87 years old), finish SHS earlier, have greater probability to be born in a rural area and went to a SHS with a larger network of migrants. Migrants also have a higher probability to have worked during the last seven days before the interview. Migrants have a lesser probability to work in agriculture, and worked more days and more hours per week than non-migrants. Table 2.3 shows that migrants have a higher probability to live in Greater Accra region than non-migrants. There is no significant difference between migrants and non-migrants in the probability to live in Central.

Figure 2.2 presents in- and out-migration by region. We see that the two regions that attract the largest number of educated youth are Greater Accra and Ashanti, and that Volta and Eastern generate the most migrants. Some regions attract very few educated youth ; Upper East, Volta, Upper West, and Eastern each receive fewer than 2% of migrants. However, there is no clear relationship between in-migration and out-migration (Figure 2.3). Greater Accra seems clearly to be an outlier, but we cannot delete it from our sample as it is an actual choice of destination for potential migrants in Ghana. We cannot pretend this choice does not exist. So we keep Greater Accra, being aware that it is an outlier. [19]

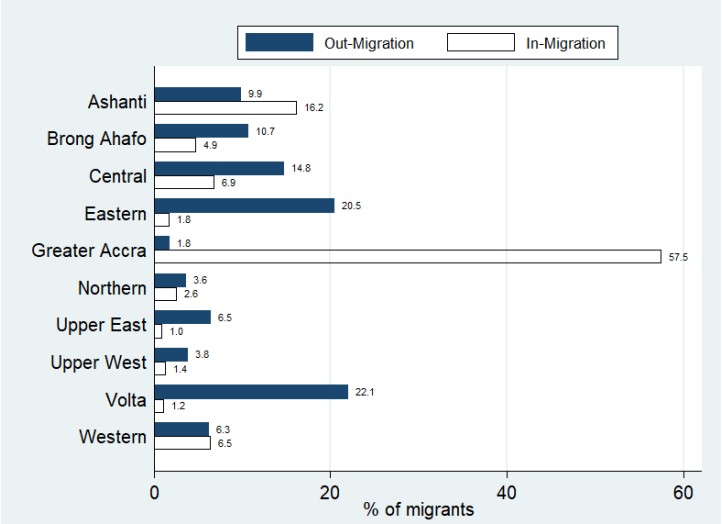
2.3.2 Average Income and Unemployment of the Ghanaians

In this chapter, we study if the respondent's expectations of income in a particular region are linked to the decision to migrate into this region. Thus, we need to know the amount each respondent expects to earn in each region of the country. However, the survey does not capture the subjective expectations of income by location. Therefore, as it is usual in the literature, we use the actual earnings by location as a proxy of the regional expected incomes. This method lies on the assumption that the respondents have a perfect information regarding the observed income in each region. Moreover, by taking the average income by region, the measure of expected incomes is less accurate than if we had the exact expected income of each respondent. We discuss in more details the limit of this method in Chapter 4 and we overcome these disadvantages by using the subjective expectations of the respondent instead of the observed incomes.

In order to analyze the link between migration decisions and average income, we need a credible source of data on average income. The sample size by region in the primary database is too small to be able to compute representative average income by region. Moreover, we want to capture the effect of general equilibrium that might occur if educated individuals migrate to higher incomes. A first effect of general equilibrium might be positive on income as educated individuals move and claim for higher incomes. But a negative effect of general equilibrium might occur as there are

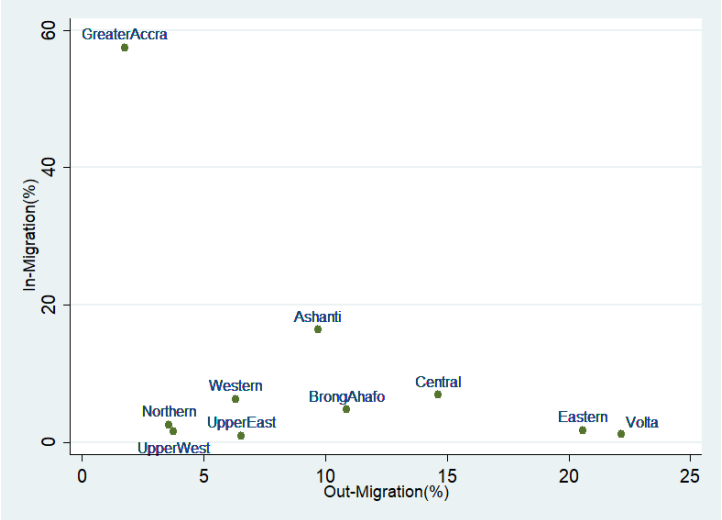
19. This is the reason why we replicate the Figure 2.6 without Greater Accra in Appendix (Figure A.1). The relationship between expected monthly income and net migration rate by region is still positive and statistically significant.

FIGURE 2.2 Migrants' Origin and Destination Regions



Note : The sum of all the 'in-migration' flows is 100% ; The sum of all the 'out-migration' flows is 100%. Anne Duplantier (2021)

FIGURE 2.3 Relationship Between In- and Out-Migration



Note : Relationship no statistically significant at 5% (p-value = 0.233). Anne Duplantier (2021)

more individuals on the labor market, which increases competition for employment and decreases incomes. The primary database might not allow the capture of these effects. Using the average incomes from a database with a larger sample size and a better representativeness would seemingly allow us to overcome these limitations. We extract the regional average incomes from the sixth round of the Ghana Living Standards Survey (GLSS6). This database was collected by the Ghana Statistical Service with support from the World Bank in 2012-13. [20]

We identify individuals who were working during the week prior to the interview. The definition of work includes any work for pay, profit, family gain or the production of anything for barter or home, and the associate pay or profit from these activities. We then calculate individual total monthly income in Ghanaian cedis (GHS). Both monthly wages and monthly in-kind payments [21] are computed by taking into account the unit of pay mentioned by the respondent, and the two amounts are added together. This measure of income accounts for any remuneration, including bonuses, commissions, allowances or tips received for work (employment, agricultural labour and self-employment) as well as the value of goods and services potentially provided as in-kind benefits. We use this income to compute regional incomes which are used as a proxy for the regional income expectations of individuals interviewed in the 'Go Transition' survey previously described. Indeed, individuals in the 'Go Transition' sample made their migration decisions prior to April 2012 [22] (when they were interviewed), whereas income data from GLSS6 were collected between October 2012 and October 2013, over the period of one year. This time gap implies that individuals who migrated in the 'Go Transition' database could not know the average incomes captured in the GLSS6 database, so they can be used as expectations of incomes.

We use the Consumer Price Index deflator 2013 from the World Bank's World Development Indicators (WDI) to deflate incomes. In this way, we take into account inflation between 2012 [23] and 2013 and express real individual total monthly income at 2012 values. We also divide real individual total monthly income by a spatial and temporal deflator to account for difference in the costs of living within country and for price difference between the month of interview. This allows us to compare average incomes in each region after taking into account price differences. Finally, we calculate the expected average income (real total monthly income) for each region by multiplying income by the probability of employment as described previously. As our sample consists of individuals who graduated from SHS and are in the majority aged between 18 and 28, [24] we compute the expected

20. According to the GLSS6 User Manual, this nationally and regionally representative sample consists of 18 000 households from both rural and urban areas. The sampling method used in this data collection is a two-stage stratified design. The primary sampling units have been distributed into the ten regions using probability proportional to population size. The GLSS6 response rate is 93.2% with 16 772 households successfully interviewed.

21. The value of the goods and services provided as a payment for the work.

22. Or prior to January 2012 for the wave of 2011.

23. 106 individuals have been interviewed from September 2011 to January 2012, but we use the same deflator.

24. 99% of our sample is aged between 18 and 28, with a minimum of 17 and a maximum of 38.

average income for several sub-samples : the sub-sample with the same level of education (SHS), the sub-sample of youth aged 18 to 28, the sub-sample of youth with SHS, and gender. We use the employment rates associated with each sub-group.

Following the International Labor Organization's definition, we consider as unemployed every individual who declares having no work in the seven last days, having been available to work and having made an effort to find work. We use population weight [25] in the calculation of average income and unemployment rate. Once the different average expected incomes and unemployment rates are built by region using the GLSS6 database, we merge these variables of income and unemployment to the 'Go Transition' database. We associate to every respondent the same regional means. This allows us to build a variable taking the value of the average expected income of the region of birth, SHS or current region for each respondent. Descriptive statistics for income and employment by region are presented in Tables 2.4-2.6.

TABLEAU 2.4 Migration, Average Income and Unemployment

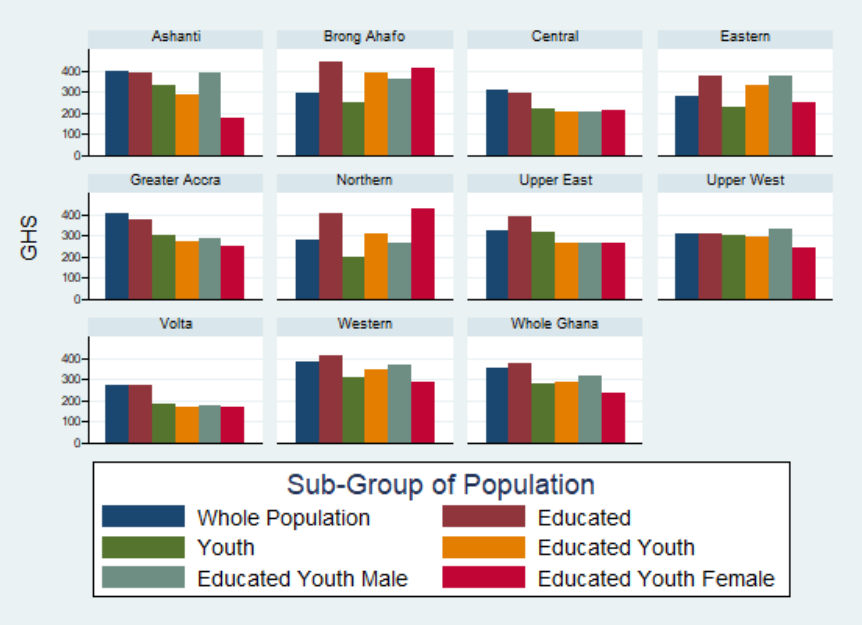
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Region	Migration in (%)	Migration out (%)	Income	Income w. SHS	Youth income	Youth income w. SHS	Male youth income w. SHS	Fem. youth income w. SHS	Unemp. rate (%)
Ashanti	16.14	9.76	410.13	403.12	344.74	294.74	402.83	180.31	2.04
			(493.06)	(473.20)	(386.29)	(355.17)	(453.15)	(136.75)	
Brong Ahafo	4.78	10.56	302.36	453.87	260.19	400.30	373.61	423.17	2.38
			(460.40)	(678.05)	(480.68)	(666.92)	(725.36)	(629.40)	
Central	6.97	14.74	319.20	305.22	227.16	214.37	211.66	216.37	2.98
			(410.92)	(404.73)	(297.38)	(278.08)	(218.62)	(319.01)	
Eastern	1.79	20.72	287.22	388.32	237.17	342.70	390.27	256.33	2.02
			(403.07)	(413.80)	(283.84)	(379.64)	(386.68)	(359.53)	
Greater Accra	57.77	1.79	417.67	384.43	312.84	281.29	297.50	257.82	3.96
			(436.84)	(397.63)	(372.68)	(361.33)	(427.33)	(236.80)	
Northern	2.59	3.59	289.21	415.50	201.68	316.63	270.87	444.02^{+}	1.84
			(414.65)	(506.27)	(246.25)	(475.02)	(264.96)	(832.90)	
Upper East	1	6.37	332.36	404.54	327.52	275.79	276.82	275.36	7.97
			(329.69)	(314.16)	(252.60)	(184.41)	(277.49)	(135.73)	
Upper West	1.39	3.78	321.39	321.13	314.30	300.44	338.40	250.39^{+}	1.24
			(376.61)	(287.22)	(287.90)	(319.58)	(320.71)	(338.59)	
Volta	1.2	22.31	284.13	281.24	191.94	176.30	179.48	172.63	2.1
			(400.82)	(335.46)	(239.42)	(178.00)	(182.72)	(176.46)	
Western	6.37	6.37	395.42	423.82	321.26	358.09	380.03	296.55	2.65
			(489.15)	(441.51)	(372.01)	(371.70)	(322.54)	(487.14)	
Ghana			361.23	384.35	287.70	292.80	328.60	245.89	2.72
			(449.39)	(437.79)	(356.72)	(377.47)	(413.14)	(319.61)	
Observations	502	502	14157	1432	3246	596	345	251	14559

Note : + Fewer than ten observations used to compute mean. Standard deviations are in parentheses. The variables 'Migration in' and 'Migration out' are from the authors' 'Go Transition' database. All other variables are calculated using the GLSS6 database. Mean income is deflated and expressed in GHS. Anne Duplantier (2021)

Table 2.4 and Figure 2.4 present the average incomes for different sub-groups of the population by region as well as the percentage of individuals from, and moving into, each region. The regions attracting the most migrants (Greater Accra and the Ashanti regions) are those with the highest average monthly incomes ; 418 and 410 GHS respectively. Conversely, there are more migrants from the Volta and Eastern regions than from any other region (between 16% and 22%), whether measuring migration from SHS or from birth. The average incomes in these regions are the lowest,

25. As less populated regions were oversampled in GLSS6, weights were computed by the Ghana Statistical Service as the inverse of the probability of selection of the household, computed at the enumeration area level.

FIGURE 2.4 Average Income by Sub-group of Population



Anne Duplantier (2021)

284 and 287 GHS respectively, and thus below the national average of 361 GHS.

There is no clear relationship between average income and that of the educated sub-sample. In some regions, such as Brong Ahafo, Eastern and Northern, the average income of individuals with an SHS diploma is above the average, but in others (such as Ashanti, Central and Greater Accra), it is below the regional average. One possible explanation is the average level of education by region. In the whole country the average level of education is around 5.4 years, while it is 6.1 and 7.3 years in Ashanti and Greater Accra, respectively, and 5.2 and 4.3 year in Brong Ahafo and Northern, respectively. [26] These differences in average education level may explain the differences in the average income pattern. In regions such as Ashanti and Greater Accra, where people are more educated, the competition is stronger for educated incomes than in regions such as Brong Ahafo and Northern, where education level is low. The average income of individuals between 18 and 28 years of age is below the regional average for every region.

Finally, the last column of Table 2.4 provides the unemployment rates in each region. Indeed, if youth are attracted by high average income, rational individuals will also consider the unemployment rate, which may create a disincentive to migration. We observe that three regions have an unemployment rate higher than the national average of 2.72% : Upper East with 7.97%, Greater Accra with 3.96% and Central with 2.98%. We observe that, even though the unemployment rate

26. Numbers computed from the GLSS6 database.

in Accra is above the national mean, it still attracts many migrants in the sample.

TABLEAU 2.5 Unemployment Rate by Region and Sub-Group

Region	(1) Unemp. rate	(2) Unemp. rate w. SHS	(3) Youth unemp. rate	(4) Youth unemp. rate w. SHS	(5) Male youth unemp. rate w. SHS	(6) Fem. youth unemp. rate w. SHS
Ashanti	2.04	4.89	4.96	7.15	9.47	4.56
Brong Ahafo	2.38	8.73	7.72	16	20.2	12.03
Central	2.98	8.4	8.62	13.18	22.42	4.8
Eastern	2.02	5.19	6.84	8.86	7.04	11.99
Greater Accra	3.96	5.22	10.05	9.05	6.25	12.82
Northern	1.84	2.97	3.78	6.06	4.81	9.39
Upper East	7.97	14.63	13.26	22.85	10.78	26.94
Upper West	1.24	1.41	3.79	4.06	0	8.94^{+}
Volta	2.1	6.25	5.31	11.04	15.49	5.28
Western	2.65	5.4	7.74	10.38	6.65	19.41
Ghana	2.72	5.65	7.31	9.68	9.41	10.04
Observations	14559	1541	3502	680	390	290

Note : + Number of observations fewer than ten. Anne Duplantier (2021)

Table 2.5 reports the unemployment rate by region and sub-group. The youth unemployment rate exceeds that of the overall population and an SHS diploma does not always improve the probability of finding a job. In addition, while the average national unemployment rate is 2.72% for all individuals, it is 9.68% for youth with an SHS degree. This rate is slightly lower for men than for women.

If the decision to migrate depends on earnings expectation, it is a function of average income and the probability to find a job, i.e. unemployment rate in the destination region. It is therefore important to understand the relationship between average income and unemployment rate in each region. In Figure 2.5, we present the relationship between average income and unemployment rate for the total population in each region. We observe a positive relationship between mean income and the unemployment rate, but the relationship is not statistically significant. [27]

The probability of employment is taken into account directly in Table 2.6, where we present expected mean incomes, computed as the product between average income and average unemployment rate, by region for the same sub-group. Regional differences in expected average income from Table 2.6 and average income from Table 2.4 are similar. The remainder of the analysis uses the expected average income.

Figure 2.6 shows a strong positive and statistically significant correlation between expected average income and net migration to each region. From the 'Go Transition' database presented previously,

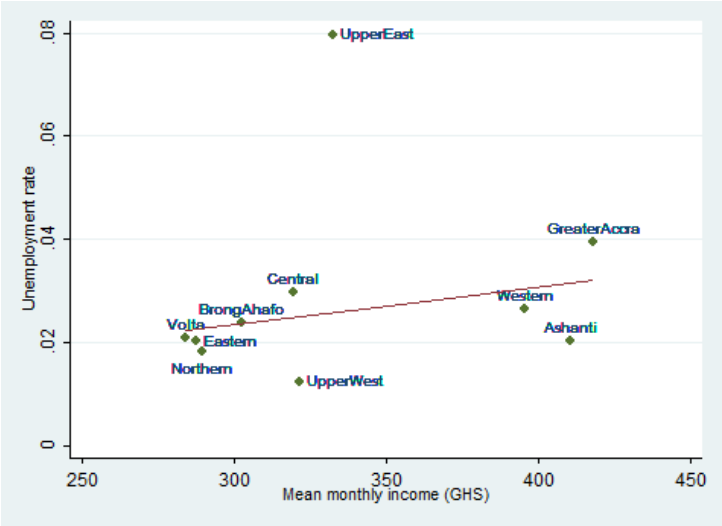
27. The relationship does not change when we drop the outlier Upper East.

TABLEAU 2.6 Expected Average Income by Region and Sub-Group

Region	(1) Exp. income	(2) Exp. income w. SHS	(3) Exp. youth income	(4) Exp. youth income w. SHS	(5) Exp. male youth inc. w. SHS	(6) Exp. fem. youth inc. w. SHS
Ashanti	399.56	392.73	335.86	287.15	392.45	175.66
	(480.35)	(461.00)	(376.34)	(346.01)	(441.47)	(133.22)
Brong Ahafo	294.57	442.17	253.48	389.98	363.98	412.26
	(448.53)	(660.57)	(468.29)	(649.73)	(706.67)	(613.17)
Central	310.97	297.35	221.30	208.84	206.21	210.79
	(400.33)	(394.29)	(289.71)	(270.91)	(212.99)	(310.79)
Eastern	279.82	378.31	231.06	333.87	380.21	249.72
	(392.68)	(403.13)	(276.52)	(369.85)	(376.71)	(350.26)
Greater Accra	406.91	374.52	304.77	274.04	289.83	251.18
	(425.58)	(387.38)	(363.07)	(352.02)	(416.31)	230.69)
Northern	281.75	404.79	196.48	308.47	263.89	432.57^{+}
	(403.96)	(493.22)	(239.90)	(462.77)	(258.13)	(811.44)
Upper East	323.79	394.11	319.08	268.68	269.69	268.26
	(321.19)	(306.06)	(246.09)	(179.65)	(270.34)	(132.23)
Upper West	313.10	312.85	306.19	292.70	329.68	243.93^{+}
	(366.90)	(279.82)	(280.48)	(311.34)	(312.44)	(329.86)
Volta	276.80	273.99	187.00	171.76	174.85	168.18
	(390.48)	(326.81)	(233.25)	(173.41)	(178.01)	(171.91)
Western	385.23	412.90	312.98	348.86	370.24	288.90
	(476.54)	(430.13)	(362.42)	(362.12)	(314.22)	(474.58)
Ghana	351.92	374.45	280.29	285.25	320.13	239.55
	(437.81)	(426.50)	(347.52)	(367.74)	(402.49)	(311.38)
Observations	14157	1432	3246	596	345	251

Note : + Fewer than ten observations used to compute mean. Standard deviations are in parentheses. All variables are calculated using the GLSS6 database. Mean income is deflated and expressed in GHS. Anne Duplantier (2021)

FIGURE 2.5 Relationship Between Average Income and Unemployment Rate in Each Region

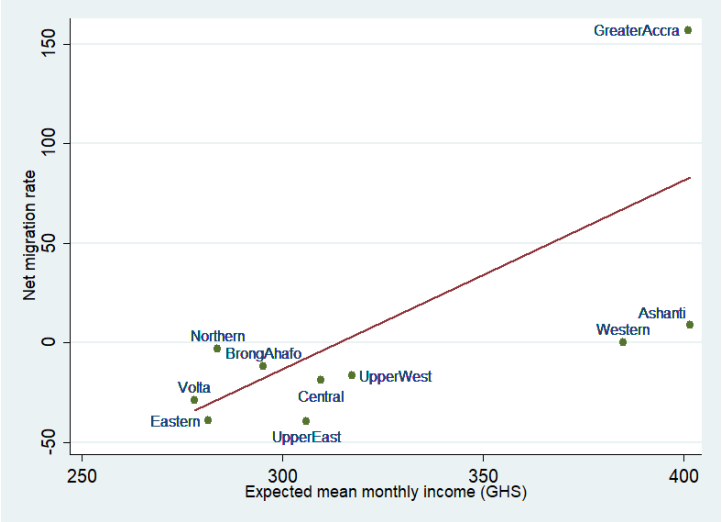


Note : Relationship non statistically significant at 5% (p-value = 0.615). The relationship does not change when we drop the outlier Upper East.

Anne Duplantier (2021)

we calculate net migration by excluding return migration and taking the difference between in-migration and out-migration by region. Finally, we divide net migration by the total population prior to migration for each region, in order to obtain net migration rate. The higher a region's expected mean income, the more it is preferred as a destination for migrants. Thus, *a priori*, there seems to be a link between expected mean income and inter-regional net migration in Ghana. Even when we exclude the Greater Accra Region (cf Figure A.1 in Appendix), the positive correlation persists and remains statistically significant.

FIGURE 2.6 Relationship Between Expected Average Income and Net Migration to Each Region



Note : Relationship positive ($\beta = 0.672$) and statistically significant at 5% (p-value = 0.033). Anne Duplantier (2021)

2.4 METHODOLOGY

We assume that Ghanaian youth can choose to live in any of the ten regions of Ghana. Our dependent variable is a categorical variable for the location of residence of the respondent and is equal to one of the ten regions of Ghana. As such, we consider multinomial models for our estimation. In our situation, the values of the possible alternatives (the ten regions) are not ordered in any particular order. Therefore we estimate a multinomial unordered model. The category of multinomial unordered model includes different types of model that we present briefly below in order to discriminate the more appropriate one. In all of these models, variables are discriminated into case-specific variables and alternative-specific variables. In our scenario, the case-specific variables refer to individual-specific variables (for example, age, sex or parental education) and the alternative-specific variables refer to region-specific variables (for example, average income).

The multinomial logit model (MNL) is the simplest model appropriate for databases which have only case-specific variables. With these types of databases, independent variables are only observed for the chosen alternative and not for the other alternatives. The MNL model is a discrete choice model including variables specific to each case but in which the attributes of the alternatives remain constant. This is the estimation technique used by De Vreyer et al. (2009) when analyzing migration between the countries of the West African Economic and Monetary Union. Alternatively, when a richer database with information regarding all alternatives, not only the chosen one, is available, the conditional logit model developed by McFadden (1974) is appropriate.

These two situations can be combined when we have access to both information - cases and al-

ternatives. The unordered multinomial model used in this case is the mixed logit model, which accommodates variables specific to both individuals and regions. This is the model we use here, as our database includes individual characteristics and the attributes of the regions. Boyd and Mellman (1980) and Cardell and Dunbar (1980) are the first to have used this method when modeling automobile demand. The key underlying assumption is that rational individuals consider the characteristics of each option and then choose the one that maximizes their utility. This method is also known as 'the random coefficients' multinomial logit model or mixed multinomial logit model (MMNL). McFadden and Train (2000) established that under weak conditions, MMNL models are random utility maximization models that are frequently used in the literature on international migration.

In the mixed logit model, the expected utility of a choice may depend on the characteristics of both the alternatives and of the individuals making the choice, as well as on variables specific to the combination of the person and the alternative. We argue that this is likely the case for migration decisions. Moreover, the assumption of the independence of irrelevant alternatives (IIA) in MNLs, is relaxed in the mixed logit model to allow for the existence of alternatives with similarities (for example, two regions with the same main ethnic group). Indeed, when alternatives have unobserved attributes in common, this assumption is rejected. Thus, with a mixed logit model a change in one alternative may not have a proportional impact on the probability of choosing the other alternatives (the ratios of probabilities are affected). This is not the case with standard MNLs, where the probability ratios are constant.

The advantage of the mixed logit model is that it overcomes the limitations of the standard logit model. Train (2009) notes that the mixed logit model allows for random taste variation, meaning that differences in taste can be linked to observed characteristics, whereas standard logit models control only for systematic taste variation. The second limitation mentioned above is that logit models assume the independence of irrelevant alternatives. Mixed logit models allow for unrestricted substitution patterns and relax the IIA assumption. The final limitation of standard logit models suggested by Train (2009) is that they cannot capture situations where unobserved factors are correlated over time, whereas mixed logit models allow for correlations in unobserved factors over time.

Following Wooldridge (2002), let the utility of individual i from choosing alternative j from among the set of alternatives J be :

$$U_{ij} = \delta_i V_{ij} + \epsilon_{ij} \tag{2.1}$$

where V_{ij} is a vector of explanatory observed non-stochastic variables, including the socio-economic characteristics of individual $i = 1, \ldots, I$ and the attributes of alternative $j = 1, \ldots, J$. δ_i and ϵ_{ij} are not observed and treated as stochastic elements. The logit model imposes that ϵ_{ij} is

a random term that is independently and identically distribued (iid) extreme value. Cameron and Trivedi (2010) call this type of model the “additive random-utility model".

Hensher and Greene (2003) explain that the assumption of iid errors is restrictive as it does not allow for the error components of different alternatives to be correlated. The mixed logit model partitions the stochastic component additively into two parts. One part is correlated over alternatives and heteroskedastic, and the other part is iid over alternatives and individuals. When decomposing ϵ_{ij} from Equation 2.1, we obtain :

$$U_{ij} = \delta_i V_{ij} + [\eta_{ij} + \mu_{ij}] \tag{2.2}$$

where η_{ij} is a random term with zero mean with a normal distribution over individuals and alternatives and correlated across alternatives. In the standard logit model, the absence of correlation in utility over alternatives gives rise to the IIA property and its restrictive substitution patterns. μ_{ij} is a random term with zero mean that is iid over alternatives.

Cameron and Trivedi (2010) decompose the vector of explanatory variables V_{ij} from Equation 2.1 in two parts :

$$V_{ij} = X'_{ij}\beta + Z'_i\gamma_j \tag{2.3}$$

where X_{ij} is a vector composed of x_{ij}, the alternative-specific regressors, i.e. the independent variables that vary over alternatives and possibly vary across individuals. β is a set of coefficients that do not vary over individuals or alternatives, Z_i is a vector composed of z_i that are the case-specific regressors, i.e. the independent variables that vary over individuals but not over alternatives, and γ_j is a set of coefficients that vary over alternatives.

Therefore, we can rewrite Equation 2.1 as follow :

$$U_{ij} = X'_{ij}\beta + Z'_i\gamma_j + \epsilon_{ij} \tag{2.4}$$

with ϵ_{ij} the unobserved error term.

Let y_i be the choice of individual i maximizing utility :

$$y_i = argmax(U_{i0}, U_{i2}, \ldots, U_{iJ}). \tag{2.5}$$

We observe the outcome $y_i = j$ (i.e. individual i chooses alternative j) if the utility of individual i when choosing alternative j is the highest utility among the alternatives. It follows that the probability of individual i selecting alternative j is modeled as the probability of U_{ij} being greater than

all other U_{ik} , $k \neq j$ (Cameron and Trivedi, 2010) :

$$\begin{aligned} Pr(y_i = j) &= Pr(U_{ij} \geq U_{ik}), && \text{for all } \ k \neq j \\ &= Pr(U_{ij} - U_{ik} \geq 0), && \text{for all } \ k \neq j \\ &= Pr(V_{ij} + \epsilon_{ij} - V_{ik} - \epsilon_{ik} \geq 0), && \text{for all } \ k \neq j \\ &= Pr(\epsilon_{ik} - \epsilon_{ij} \leq V_{ij} - V_{ik}), && \text{for all } \ k \neq j \end{aligned} \tag{2.6}$$

In a logit model, we assume that $\epsilon_{ik} - \epsilon_{ij}$ follows a logistic distribution, and that ϵ_{ik} and ϵ_{ij} themselves are extreme value random terms.

If we consider F as the logistic distribution function with mean 0 and scale parameter 1, we have :

$$\begin{aligned} Pr(y_i = j) &= F(V_{ij} - V_{ik}), && \text{for all } \ k \neq j \\ &= F((X'_{ij}\beta + Z'_i\gamma_j) - (X'_{ik}\beta + Z'_i\gamma_k)), && \text{for all } \ k \neq j \end{aligned} \tag{2.7}$$

As shown in Train (2009), the mixed logit choice probabilities are :

$$Pr(y_i = j|X_i) = \frac{exp(X'_{ij}\beta + Z'_i\gamma_j)}{\sum_{k=0}^{J} exp(X'_{ik}\beta + Z'_i\gamma_k)}, \tag{2.8}$$

where X_i is a vector containing x_{ij} for all values of j, i.e. $j = 1, \ldots, J$; and x_{ik} is the vector of observed non-stochastic variables for individual i and alternative $k = 1, \ldots, J$.

The impact of a change in x_{ij} is

$$\frac{\partial Pr_{ij}}{\partial x_{ij}} = \begin{cases} Pr_{ij}(1 - Pr_{ij})\beta & \text{if } \ j = k \\ -Pr_{ij}Pr_{ik}\beta & \text{if } \ j \neq k \end{cases} \tag{2.9}$$

If $\beta > 0$ the direct effect of a change in x_{ij} is positive since $Pr_{ij}(1 - Pr_{ij})\beta > 0$ and the cross effect is negative because $-Pr_{ij}Pr_{ik}\beta < 0$. A positive coefficient indicates that, if variable x_{ij} (for an alternative j) increases, there is a greater probability of this alternative being chosen (direct effect) and a lower probability of the other alternatives being chosen (cross effect).

We adapt this mixed logit model to the choice of region of residence made by youth who previously attended SHS in Ghana. Thus, each Ghanaian youth i chooses their region of residence j from among the ten regions of Ghana. The first estimated model is linearly represented and specified as follows :

$$y_{ij} = \beta_{0j} + \beta_{ratio} Ratio_{ij} + \gamma_{1j} Boarding_i + \gamma_{2j} Ability_i + \gamma_{3j} Age_i + \gamma_{4j} Age_i^2 + \gamma_{5j} Rural_i + \gamma_{6j} Male_i + \gamma_{7j} YearDiploma_i + \gamma_{8j} Network_i + \gamma_{9j} DistanceAccra_i + \epsilon_{ij} \tag{2.10}$$

with

$$Ratio_{ij} = \frac{Origin\ region's\ expected\ income}{Destination\ region's\ expected\ income} \tag{2.11}$$

where β_{0j} is the alternative specific constant (ASC) for region j; β_{ratio} is the expected income ratio sensitivity parameter; $Ratio_{ij}$ is the ratio of origin-region expected income to destination-region for individual i if they were to live in region j; $Boarding_i$, $Ability_i$, Age_i, Age_i^2, $Rural_i$, $Male_i$, $YearDiploma_i$, $Network_i$, and $DistanceAccra_i$ are the socio-economic variables related to migration and described in Table 1; and ϵ_{ij} is the iid error term for individuals and alternatives that can be decomposed in $\eta_{ij} + \mu_{ij}$ as shown in Equation 2.2.

We think that being in a boarding SHS is related with a higher probability to migrate as the student has already left the family house. Therefore, their probability to migrate again is higher. As we saw previously in the literature, the relationship between ability or education of a respondent and their probability to migrate is uncertain. A respondent with high abilities living in a rural area will have a higher probability to move toward an urban area where there is a higher proportion of educated jobs. At the same time, we can imagine that a respondent with low abilities living in a rural area will migrate to an urban area because the demand of (educated or not) labor is higher. Regarding the age, we think that there is a positive relationship between age and migration, as the ability to migrate is higher if the individual is older and more independent. However, as we think that the relationship between age and migration is an inverted U shape, the age squared is included. Indeed, after a certain age the probability to migrate decreases as the individual gets a family (partner and child) and more settled in their life.

Being born in a rural area can be a factor increasing the probability to migrate, as the student has already moved to go to SHS. We find in the literature review that in general men have a higher probability to migrate than women. Regarding the number of years after diploma, we think that the more the student waits without opportunity after SHS, the higher the probability will be of them migrating. The network of migrants will have a positive effect on the probability to migrate. The more the student has migrants in their SHS, the easier it will be for them to migrate, because the network brings information and material help. Having migrants in the SHS also increases the probability to migrate as it places the idea of migration and shows that it is accessible. Finally, the distance to Accra is negatively related to the migration toward Accra, but positively to neighbor region, as distance might be an obstacle to migration (in terms of cost, for example).

The mixed logit model is estimated with ten observations for every individual, representing each of the ten potential destination regions in Ghana : for 1878 individuals in the sample, the database includes 18 780 observations. [28] Thus, the variable for the income ratio corresponds to the ratio between the income in the region of SHS (region of origin) and the income in each of the ten regions of Ghana (the potential destination regions). Hence, the variable $Ratio_{ij}$ varies across individuals (SHS regions) and alternatives (regions in our case).

Following Harris and Todaro (1970) and Lewis (1954), who show that individuals are attracted to income differentials between rural and urban areas, we expect the coefficient β_{ratio} to be negative. *Ceteris paribus*, when the expected income in a destination region increases, the probability of migrating to that region also increases. Similarly, if the income in the region of origin increases, the likelihood of leaving that region decreases. With regard to the coefficients of the variables specific to the individuals γ_{nj} with $n = 1, \ldots, 9$ representing the different regressors included in the model, they are interpreted as the parameters of a binary logit with respect to the baseline category. If γ_{nj} is negative, an increase in the value of explanatory variable n decreases the probability of alternative j being chosen over alternative k.

2.5 RESULTS

2.5.1 Main Estimates

Our primary results are presented in Tables 2.7-2.9. The model, with expected income ratios between SHS region and destination region, is presented in Table 2.7. The coefficient $\hat{\beta}_{ratio}$ of the alternative-specific regressors (the ten regions of Ghana), i.e. the ratio of the two regions' expected incomes, is presented in the first row. The coefficients, $\hat{\gamma}_{nj}$, of individual-specific variables $n = 1, \ldots, 9$, and of the constant, $\hat{\beta}_{0j}$, are also presented in the table.

The results of each region are in comparison with the Central region. [29] We chose the Central region as the benchmark for three main reasons. First, the Central region is economically close to the average levels of income and unemployment in Ghana. Second, we saw in Table 2.3 that there is no significant difference between migrants and non-migrants in the probability to live in this region. Finally, there is a substantial amount of in- and out-migration in this region. As errors can be correlated within SHS, we cluster at SHS level in every regression.

We find that $\hat{\beta}_{ratio}$ is negative, indicating that when expected incomes increase in the region of origin and they do not vary in the potential destination regions (i.e. the relative expected incomes), the

28. The original database is transformed : we expand the data and replicated the observations in order to obtain ten observations for each individual. Each individual has an observation for every destination region.

29. In Equations 3.10 and 3.12, $\gamma_j = 0$ for $j = Central$.

TABLEAU 2.7 Expected Income Ratios Between SHS and Destination Regions

Alternative-specific variables									
Ratio of Exp. Inc.				-37.348***					
GLSS6 population				(2.858)					
Individual-specific variables									
Baseline region :		Brong		Greater		Upper	Upper		
Central	Ashanti	Ahafo	Eastern	Accra	Northern	East	West	Volta	Western
Boarding	0.349	-0.427	0.114	0.296	-0.413	0.380	0.824	-0.636*	-0.227
	(0.300)	(0.363)	(0.335)	(0.329)	(0.385)	(0.566)	(0.644)	(0.355)	(0.340)
Cognitive abilities	-0.001	-0.002	0.001	0.000	0.005	0.011	0.001	-0.004	0.000
	(0.006)	(0.007)	(0.006)	(0.006)	(0.007)	(0.010)	(0.012)	(0.005)	(0.006)
Age	-0.542	0.566	1.775**	-0.354	0.419	0.036	1.103	-0.286	0.527
	(0.404)	(0.526)	(0.739)	(0.438)	(0.676)	(0.956)	(0.884)	(0.463)	(0.514)
Age squared	0.014	-0.011	-0.038**	0.009	-0.006	0.001	-0.018	0.009	-0.011
	(0.009)	(0.012)	(0.017)	(0.009)	(0.014)	(0.021)	(0.020)	(0.010)	(0.012)
Born rural areas	-0.934***	-0.255	0.499*	-0.748***	-0.404	0.111	-0.714	0.428*	-0.089
	(0.265)	(0.302)	(0.262)	(0.252)	(0.351)	(0.452)	(0.473)	(0.252)	(0.261)
Male	0.299	0.407	0.244	0.141	0.643	0.620	0.321	0.137	0.798**
	(0.287)	(0.379)	(0.284)	(0.284)	(0.432)	(0.479)	(0.928)	(0.321)	(0.316)
Years since diploma	-0.006	-0.052	0.083	0.090	-0.094	0.020	-0.116	-0.067	0.055
	(0.090)	(0.097)	(0.085)	(0.085)	(0.106)	(0.151)	(0.150)	(0.081)	(0.095)
Percentage migrants	-3.125*	-7.151***	1.861	4.911***	-14.678***	1.329	-4.677	-1.783	-0.720
per SHS	(1.821)	(2.663)	(1.907)	(1.787)	(3.335)	(2.152)	(4.113)	(2.119)	(2.032)
Distance SHS-Accra	2.194***	3.312***	-0.808***	-0.330	5.934***	4.734***	6.183***	0.959***	1.581***
(log)	(0.336)	(0.431)	(0.302)	(0.258)	(0.804)	(1.509)	(1.716)	(0.331)	(0.333)
Constant	-13.461**	-20.695***	-14.828**	-3.785	-33.502***	-29.909***	-51.498***	2.199	-22.129***
	(5.355)	(6.501)	(8.744)	(5.514)	(9.172)	(9.227)	(12.265)	(6.017)	(6.034)
Observations				18780					
Individuals				1878					
Mcfadden-R2				0.068					

Note : Robust standard errors are in parentheses and there are cluster at SHS level. Baseline region is Central. *** p<0.01, ** p<0.05, * p<0.1. As explained in the methodology section, the number of observations is ten times the number of individuals. Anne Duplantier (2021)

probability of migrating falls significantly. This means that young educated Ghanaians are attracted by high relative expected incomes in other destination regions of Ghana. Another potential interpretation is that they are attracted by a region where there is a large supply of leisure and recreation activities. Indeed, it is very likely that a region where earnings are high is more able to develop and offer recreational activities. As we do not have a way to capture these variables, we cannot differentiate both effects (high income and leisure).

Furthermore, individuals born in a rural area are less likely to choose the Ashanti and Greater Accra regions compared with the Central region, but are more likely to choose to migrate to the Eastern region than the Central region. One possible explanation is that respondents coming from rural areas are more attracted by rural regions such as Eastern. Men are more likely to migrate to the Western region than to the Central region (no significant effect for the other regions). We cannot interpret the magnitudes of effect in this model, so we will look at marginal effects further (cf Tables 2.12 and 2.13).

In terms of the characteristics of the individual's educational institution, the distance between the school and the capital of Ghana affects the choice of destination region. Thus, the greater the distance between the SHS and Accra, the higher the likelihood the migrants will choose the Ashanti, Brong Ahafo, Northern, Upper East, Upper West and Western regions over the Central region. The Central region is one of the regions that borders Greater Accra. If the SHS is far from Accra, the capital of Ghana where economic opportunities are concentrated, individuals will tend to move to

regions that are closer to home rather than to the Central region. However, an increased distance from Accra also reduces the probability of moving to Greater Accra or to the Eastern region relative to the Central region. Therefore, unsurprisingly, individuals who are far from Accra are less prone to choose to migrate to Greater Accra and to Eastern, a neighboring region, than to the Central region, which may be closer to their SHS.

TABLEAU 2.8 Expected Income Ratios Between SHS and Destination Regions - Youth with SHS

Alternative-specific variables									
Ratio of Exp. Inc.				-9.099***					
Youth w. SHS				(0.868)					
Individual-specific variables									
Baseline region :		Brong		Greater		Upper	Upper		
Central	Ashanti	Ahafo	Eastern	Accra	Northern	East	West	Volta	Western
Boarding	-0.068	-0.544	-0.244	-0.085	-0.475	0.208	0.587	-0.675*	-0.638*
	(0.286)	(0.371)	(0.348)	(0.307)	(0.402)	(0.570)	(0.616)	(0.356)	(0.353)
Cognitive abilities	0.002	0.001	0.004	0.003	0.007	0.011	0.004	-0.006	0.004
	(0.006)	(0.007)	(0.007)	(0.006)	(0.007)	(0.010)	(0.012)	(0.005)	(0.006)
Age	-0.846*	-0.192	0.882	-0.715	-0.182	-0.368	0.619	-0.323	-0.028
	(0.463)	(0.508)	(0.686)	(0.460)	(0.663)	(0.929)	(0.887)	(0.515)	(0.572)
Age squared	0.020**	0.007	-0.016	0.016	0.008	0.010	-0.007	0.010	0.001
	(0.010)	(0.011)	(0.015)	(0.010)	(0.014)	(0.019)	(0.020)	(0.011)	(0.013)
Born rural area	-0.948***	-0.302	0.394	-0.747*	-0.428	0.092	-0.761	0.462*	-0.110
	(0.271)	(0.324)	(0.276)	(0.223)	(0.338)	(0.476)	(0.463)	(0.245)	(0.281)
Male	-8.967***	-0.198	-6.164***	-3.910***	2.688***	-3.399***	-5.301***	-1.392***	-5.017***
	(0.768)	(0.422)	(0.605)	(0.400)	(0.586)	(0.666)	(1.037)	(0.375)	(0.618)
Years since diploma	-0.010	-0.042	0.050	0.098	-0.084	0.034	-0.096	-0.066	0.053
	(0.093)	(0.103)	(0.095)	(0.081)	(0.114)	(0.153)	(0.152)	(0.080)	(0.101)
Percentage migrants	-6.557*	-3.977**	5.166***	1.596	-12.310***	1.814	-3.887	-1.618	-2.792
per SHS	(1.789)	(1.699)	(1.570)	(1.468)	(3.115)	(2.034)	(4.135)	(1.734)	(2.186)
Distance SHS-Accra	1.757***	2.949***	-0.509**	-0.434**	6.508***	4.196***	5.783***	1.128***	1.222***
(log)	(0.300)	(0.439)	(0.233)	(0.213)	(0.834)	(1.411)	(1.416)	(0.364)	(0.299)
Constant	3.937	-18.687***	-12.304	9.910*	-37.957***	-22.770***	-44.500***	-0.120	-7.330
	(5.607)	(6.439)	(8.076)	(5.464)	(8.823)	(8.757)	(11.539)	(6.486)	(6.531)
Observations				18780					
Individuals				1878					
Mcfadden-R2				0.072					

Note : Robust standard errors are in parentheses and there are cluster at SHS level. Baseline region is Central. *** p<0.01, ** p<0.05, * p<0.1. Anne Duplantier (2021)

In Table 2.8, we replicate the results of Table 2.7 using the expected mean income calculation for the sub-sample of youth with an SHS degree from GLSS6. We find that $\hat{\beta}_{ratio}$ remains negative, indicating that when relative expected incomes of educated youth rise in the region of SHS, the probability of leaving this region decreases for our respondents. But if we compare the scale of this coefficient to the estimation in Table 2.7 using the mean income of the full GLSS6 sample, we conclude that the relative expected incomes of the entire population of the region appears to have a larger impact than the relative expected incomes of the sub-sample of educated youth. This suggests that educated Ghanaian youth in our sample consider the income of the full population rather than their peers when deciding to where they will migrate. Alternatively, it may suggest that they consider both average incomes but the full population average income has a bigger effect. One possible explanation is that students have a long-term vision and take decisions by thinking further ahead than the near future. The estimated coefficients of the individual-specific variables are quite similar both in terms of sign and magnitude. [30]

30. There are slight differences regarding the gender though. Whereas the gender was previously not a significant characteristic, it now is in every region aside from Brong Ahafo. Being a man is related to a lower probability to migrate to almost every region than to migrate to Central.

TABLEAU 2.9 Ratio of Expected Income by Gender

	Full population		Youth with SHS	
	(1)	(2)	(3)	(4)
	Men	Women	Men	Women
Alternative-specific variables				
Ratio of Exp. Inc.	-37.919***	-37.369***	-8.917***	-12.680***
	(3.305)	(4.764)	(1.058)	(2.985)
Individual-specific variables	Yes	Yes	Yes	Yes
Observations	11380	7400	11380	7400
Individuals	1138	740	1138	740
Mcfadden-R2	0.078	0.072	0.083	0.079

Note : Robust standard errors are in parentheses and there are cluster at SHS level. *** p<0.01, ** p<0.05, * p<0.1. Anne Duplantier (2021)

Finally, we replicate these estimations but for men and women separately to investigate any differences by gender. In Table 2.9, we present the coefficient results for our primary variable of interest, the gender-specific ratio of expected income. We computed a ratio for men and women separately as we think that women will identify more to the girls' average income (and the same for men). We can legitimately wonder if the young women and men have really access to this information and use it to take a decision. Indeed, respondents are unlikely to have accurate information about the average income of women and men. However, it is common knowledge that women in all countries have on average lower incomes than men. Furthermore, it can be assumed that individuals who identify with their same-sex counterparts may have access to gender-specific income amounts even if these values are not very precise (from their entourage, friends, family). Thus, this hypothesis is certainly not perfect but seems plausible. We find that the ratio of expected income has the same effect on migration for both men and women as there is no statistically significant difference between the estimated coefficients.

2.5.2 Introducing Income Variability

We now wonder if the variability of the regional average income has also an effect on the migration decision. We add a ratio of income variability in the first model (Equation 2.10). Here is a linear representation of the model :

$$y_{ij} = \beta_{0j} + \beta_{ratioE} RatioE_{ij} + \beta_{ratioV} RatioV_{ij} + \gamma_j Controls_i + e_{ij} \tag{2.12}$$

with

$$RatioE_{ij} = \frac{Origin\ region's\ expected\ income}{Destination\ region's\ expected\ income} \tag{2.13}$$

and

$$RatioV_{ij} = \frac{Origin\ region's\ variance\ in\ income}{Destination\ region's\ variance\ in\ income} \tag{2.14}$$

where $RatioE_{ij}$ is the ratio of expected incomes between the origin-region and the destination-region for individual i living in region j ; $RatioV_{ij}$ is the ratio of the variance in expected income between the origin-region and the destination-region for individual i living in region j ; $Controls_i$ is the vector of control variables described previously.

Ceteris paribus, if the variance of incomes at destination increases, $RatioV$ decreases. The predicted sign of β_{ratioV} is expected to be ambiguous and likely to depend on both individual attitudes towards risk and their belief of where their expected income falls in the income distribution. For example, if an individual believes their expected income to be in the top 1% of the income distribution, a higher variance of incomes at destination will attract them to destination region. Conversely, if they believe their expected income is at the bottom of the income distribution, a higher variance of incomes at destination will not encourage them to migrate.

We present in Table 2.10 the results of the second model, with both the ratio of expected income and the ratio of the variance in income between the region of SHS and the region of destination. These results for the ratio of expected income are very similar to those described previously.

TABLEAU 2.10 Ratio of Expected Income and Ratio of The Variance in Incomes

	Full population			Youth with SHS		
	(1)	(2)	(3)	(4)	(5)	(6)
	Full sample	Men	Women	Full sample	Men	Women
Alternative-specific variables						
Ratio of Exp. Inc.	-28.471***	-29.558***	-27.771***	-12.736***	-12.175***	-14.200***
	(4.244)	(4.676)	(7.418)	(1.935)	(2.044)	(2.888)
Ratio of Var. in Inc.	-24.414***	-21.711***	-28.057***	-1.464***	-1.290***	-1.765***
	(3.906)	(4.526)	(5.723)	(0.222)	(0.293)	(0.405)
Individual-specific variables	Yes	Yes	Yes	Yes	Yes	Yes
Observations	18780	11380	7400	18780	11380	7400
Individuals	1878	1138	740	1878	1138	740
McFadden R2	0.072	0.079	0.080	0.070	0.082	0.079

Note : Robust standard errors are in parentheses and there are cluster at SHS level. *** p<0.01, ** p<0.05, * p<0.1. Anne Duplantier (2021)

For both expected income and variance in incomes, an increase in the ratio of the origin-region X to the destination-region Y significantly reduces the probability of migrating to region Y. In other words, when the variance of income in the region of origin increases, the ratio of the variance decreases and thus the probability of migrating increases. This negative coefficient implies that, on average, individuals in our sample are attracted to a region with a higher income variability. This is consistent with the findings of Heitmueller (2005), Chiswick (1978), Todaro (1980), Constant and Zimmermann (2006) and Jaeger et al. (2010) who show that migrants are more risk-loving than others.

The estimated coefficients of the ratios of relative expected incomes are stable in terms of significance and sign, but scales are slightly different. Regarding the whole population's relative expected incomes, scales decrease when we add the income variance ratio. However, we also note that the expected income ratio has a much greater incidence on the probability of migrating than the income variance ratio for the whole sample and men (Columns (1) and (2)). This suggests that average individuals and men pay much more attention to inter-regional relative expected income differentials than income variability differentials when migrating. Concerning women, the ratio of income variance is slightly higher than the ratio of expected income ratio, suggesting that women pay more or equally attention to extreme incomes than expected mean incomes.

Now let us estimate the predicted probabilities for each observation representing the individual's alternatives (the regions). Table 2.11 presents the different models' [31] prediction of in-migration by region (Columns (2)-(5)) and compares it with the actual migration decisions of the sample (Column (1)). In general, the prediction of the models are quite close to the reality.

TABLEAU 2.11 Predicted Probabilities of Choice of Region

Regions	(1) Sample	(2) Model 1	(3) Model 2	(4) Model 3	(5) Model 4
Ashanti	19.35	18.53	18.54	18.55	18.54
Brong Ahafo	10.15	9.98	10.01	10.02	10.02
Central	7.63	8.34	8.33	8.33	8.33
Eastern	6.23	6.27	6.27	6.27	6.27
Greater Accra	25.80	26.80	26.82	26.82	26.82
Northern	7.02	7.22	7.23	7.21	7.22
Upper East	2.01	1.58	1.57	1.56	1.57
Upper West	2.53	2.00	2.00	1.99	2.00
Volta	10.59	11.65	11.57	11.61	11.57
Western	7.71	7.63	7.65	7.64	7.65

Note : Model 1 : expected mean income from full population. Model 2 : expected mean income from educated youth. Model 3 : expected mean income and variance of expected income from full population. Model 4 : expected mean income and variance of expected income from educated youth Anne Duplantier (2021)

Finally, we estimate the marginal effects at the means in order to quantify the magnitude of the relationship between the ratios of the region of origin to the region of destination (relative expected incomes and variance in incomes) and the probability of migrating to different regions from that initially chosen (the region of SHS) at mean values of the characteristics of the individual and their SHS.

Table 2.12 presents the marginal effects of the SHS-region to destination-region ratio of expected

31. The first and more general model presented in Equation 2.5.

income. We set the choice of one alternative and then observe how the probability of choosing another region changes as we increase the ratio by one standard deviation. Making the assumption that the ratios increase of one standard deviation allows to compare the two ratios (variance and average) that are not measured the same way. Recall that the ratio of expected income equals to the average expected income in the individual's SHS region divided by the average expected income in the destination region. An increase in the ratio of expected income of 0.21 (i.e. one standard deviation) significantly reduces the probability of choosing the Ashanti region (by $0.21 * 3.703 = 0.78$) when this region was the first choice of destination and all other variables are fixed at their mean sample values. In addition, if the destination region was the Ashanti region, then a 0.21 increase in the ratio of expected income would significantly increase (by $0.21 * 1.883 = 0.39$) the probability of moving to Greater Accra. We also observe that this change in the ratio of expected income does not have a significant impact on the probability of migrating to the Northern, Upper East or Upper West regions, the three poorest regions of the country. The probabilities for the other regions vary between 0.78 and 0.39.

TABLEAU 2.12 Marginal Effects of the Ratio of Expected Income

	Chosen region									
Destination region	Ashanti	Brong Ahafo	Central	Eastern	Greater Accra	Northern	Upper East	Upper West	Volta	Western
Ashanti	-3.703***	0.317***	0.485***	0.143***	1.883***	0.023*	0.003	0.008	0.385***	0.451***
	(0.766)	(0.099)	(0.160)	(0.045)	(0.441)	(0.013)	(0.003)	(0.008)	(0.091)	(0.126)
Brong Ahafo	0.317***	-1.917***	0.229***	0.067***	0.889***	0.011*	0.001	0.004	0.182	0.213
	(0.099)	(0.502)	(0.080)	(0.024)	(0.261)	(0.006)	(0.001)	(0.004)	(0.045)	(0.071)
Central	0.485***	0.229***	-2.808***	0.103***	1.359***	0.017*	0.002	0.006	0.278***	0.325***
	(0.160)	(0.080)	(0.736)	(0.037)	(0.352)	(0.009)	(0.002)	(0.006)	(0.081)	(0.118)
Eastern	0.143***	0.067***	0.103***	-0.904***	0.402***	0.005*	0.8E-3	0.001	0.082***	0.096***
	(0.045)	(0.024)	(0.037)	(0.229)	(0.104)	(0.003)	(0.7E-3)	(0.001)	(0.022)	(0.032)
Greater Accra	1.883***	0.889***	1.359***	0.402***	-6.979***	0.066*	0.010	0.023	1.079***	1.263***
	(0.441)	(0.261)	(0.352)	(0.104)	(1.072)	(0.037)	(0.008)	(0.023)	(0.229)	(0.329)
Nothern	0.023*	0.011*	0.017*	0.005	0.066*	-0.153*	0.1E-3	0.3E-3	0.013*	0.015*
	(0.013)	(0.006)	(0.009)	(0.003)	(0.037)	(0.084)	(0.1E-3)	(0.3E-3	(0.007)	(0.009)
Upper East	0.003	0.001	0.002	0.8E-3	0.010982	0.1E-3	-0.025494	0.5E-4	0.002	0.002
	(0.003)	(0.001)	(0.002)	(0.7E-3)	(0.008)	(0.1E-3)	(0.019)	(0.6E-4)	(0.001)	(0.002)
Upper West	0.008	0.004	0.006	0.001	0.023	0.3E-3	0.5E-4	-0.055	0.004	0.005
	(0.008)	(0.004)	(0.006)	(0.001)	(0.023)	(0.3E-3)	(0.6E-4)	(0.054)	(0.004)	(0.005)
Volta	0.385***	0.182***	0.278***	0.082***	1.079***	0.013*	0.002	0.004	-2.287***	0.258***
	(0.091)	(0.045)	(0.081)	(0.022)	(0.229)	(0.007)8	(0.001)	(0.004)	(0.417)	(0.073)
Western	0.451***	0.213***	0.325***	0.096***	1.263***	0.015*	0.002	0.005	0.258***	-2.633***
	(0.126)	(0.071)	(0.118)	(0.032)	(0.329)	(0.009)	(0.002)	(0.005)	(0.073)	(0.633)

Note : Robust standard errors are in parentheses and there are cluster at SHS level. Baseline region is Central. *** p<0.01, ** p<0.05, * p<0.1. Anne Duplantier (2021)

The marginal effects of the ratio of the SHS-region to the destination-region income variance are presented in Table 2.13. Recall that the ratio of the income variances equals the variance of income in the individual's SHS region divided by the variance of income in the destination region. An increase in the variances ratio of 0.35 (i.e. one standard deviation) significantly reduces the probability of choosing the Ashanti region (by $0.35 * 3.175 = 1.11$) when the Ashanti region is the destination choice and all other variables are fixed at their mean sample values. In addition, if the destination region was the Ashanti region, an increase of one standard deviation of the ratio of variances significantly increases the probability of moving to Greater Accra (by $0.35 * 1.614 = 0.56$). Thus, we observe that the marginal effects of the ratio of expected income are higher than the

marginal effects of the ratio of variances. But, once we take into account the standard deviation differences, the impact of the variance of income on the probability of choosing one region over another as the destination region is more pronounced than the impact of the expected income ratio.

TABLEAU 2.13 Marginal Effects of the Ratio of the Variance in Income

	Chosen region									
Destination region	Ashanti	Brong Ahafo	Central	Eastern	Greater Accra	Northern	Upper East	Upper West	Volta	Western
Ashanti	-3.175***	0.272***	0.416***	0.123***	1.614***	0.020*	0.003	0.007	0.330***	0.386***
	(0.562)	(0.078)	(0.117)	(0.038)	(0.368)	(0.011)	(0.002)	(0.007)	(0.086)	(0.069)
Brong Ahafo	.272***	-1.644***	0.196***	0.058***	0.762***	0.009*	0.001	0.003	0.156***	0.182***
	(0.078)	(0.439)	(0.065)	(0.022)	(0.242)	(0.005)	(0.001)	(0.003)	(0.047)	(0.052)
Central	0.416***	0.196***	-2.408***	0.088***	1.165***	0.014*	0.002	0.005	0.238***	0.279***
	(0.117)	(0.065)	(0.582)	(0.032)	(0.300)	(0.008)	(0.001)	(0.005)	(0.075)	(0.083)
Eastern	0.123***	0.058***	0.088***	-0.775***	0.345***	0.004*	0.7E-3	0.001	0.070***	0.082***
	(0.038)	(0.022)	(0.032)	(0.221)	(0.107)	(0.002)	(0.5E-3)	(0.001)	(0.024)	(0.026)
Greater Accra	1.614***	0.762***	1.165***	0.345***	-5.985***	0.057*	0.009	0.020	0.925***	1.083***
	(0.368)	(0.242)	(0.300)	(0.107)	(1.034)	(0.033)	(0.006)	(0.019)	(0.272)	(0.246)
Northern	0.020*	0.009*	.014*	.004	.057*	-0.132*	0.1E-4	0.2E-3	0.011*	0.013*
	(0.011)	(0.005)	(0.008)	(0.002)	(0.033)	(0.073)	(0.1E-3)	(0.3E-3)	(0.006)	(0.007)
Upper East	0.003	0.001	0.002	0.7E-3	0.009	0.E-3	-0.021	0.4E-4	0.001	0.002
	(0.002)	(0.001)	(0.001)	(0.5E-3)	(0.006)	(0.1E-3)	(0.015)	(0.5E-4)	(0.001)	(0.001)
Upper West	0.007	0.003	0.005	0.001	0.020	0.2E-3	0.4E4	-0.047	0.004	0.004
	(0.007)	(0.003)	(0.005)	(0.001)	(0.019)	(0.3E-3)	(0.5E-4)	(0.046)	(0.004)	(0.005)
Volta	0.330***	0.156***	0.238***	0.070***	0.925***	0.011*	0.001	0.004	-1.961***	0.221***
	(0.086)	(0.047)	(0.075)	(0.024)	(0.272)	(0.006)	(0.001)	(0.004)	(0.482)	(0.062)
Western	0.386***	0.182***	0.279***	0.082***	1.083***	0.013*	0.002	0.004	0.221***	-2.258***
	(0.069)	(0.052)	(0.083)	(0.026)	(0.246)	(0.007)	(0.001)	(0.005)	(0.062)	(0.412)

Note : Robust standard errors are in parentheses and there are cluster at SHS level. Baseline region is Central. *** p<0.01, ** p<0.05, * p<0.1. Anne Duplantier (2021)

2.5.3 Do Individuals take into Account Unemployment Rates ?

In this sub-section, we try to understand the role of unemployment in the decision of migrants. First, we use the ratio of mean incomes instead of the ratio of expected incomes, in determining the probability of migration. Thus, we assume that individuals do not take the unemployment rate into account when choosing to where they will move. We continue to include the ratio of the variance in income in our estimates. The results (in Table 2.14) remain similar in terms of sign, statistical significance and magnitude. The slight difference lies in the coefficient of ratios for the whole sample with the whole population. When we use the expected incomes ratio, individuals are more attracted by this than by the variance ratio. However, when using the mean incomes ratio, the relationship is reversed. The other difference comes from the scale of the coefficient of mean incomes ratio, which is larger than the coefficient of expected incomes ratio.

TABLEAU 2.14 Ratio of Mean Income and Ratio of Variance in Income

	Full population			Youth with SHS		
	(1)	(2)	(3)	(4)	(5)	(6)
	Full sample	Men	Women	Full sample	Men	Women
Alternative-specific variables						
Ratio of Mean Inc.	-22.754***	-23.462***	-22.574***	-16.669***	-15.699***	-18.907***
	(3.548)	(4.013)	(6.093)	(2.212)	(2.425)	(3.134)
Ratio of Var. in Inc.	-24.721***	-21.950***	-28.509***	-1.232***	-1.095***	-1.444***
	(4.020)	(4.667)	(5.811)	(0.222)	(0.288)	(0.415)
Individual-specific variables	Yes	Yes	Yes	Yes	Yes	Yes
Observations	18780	11380	7400	18780	11380	7400
Individuals	1878	1138	740	1878	1138	740

Note : Robust standard errors are in parentheses and there are cluster at SHS level. *** p<0.01, ** p<0.05, * p<0.1. Anne Duplantier (2021)

Finally, we include directly the ratio of unemployment rates between origin region and destination region in the regression alongside the ratio of mean income and variance in incomes. Results in Table 2.15 show that ratio of unemployment is negatively and significantly related to the decision to migrate. When the unemployment rate increases in the destination region, individuals have a higher probability to migrate into this region. This can be explained by the probability to have a high unemployment rate associated with high incomes. As people are attracted by high incomes, labor supply exceeds demand, thereby creating unemployment. Theoretically, there is a threshold where unemployment discourages people from migrating, as even if the average income is high, the probability to find a job is too low.

But in this case, unemployment does not seem to prevent educated Ghanians from migrating. People migrate even if the unemployment rate is high in the destination region. This suggests that average income are more important for migrants than unemployment and the variance of income. This is consistent with what is found regarding the marginal effects (marginal effects of average income are higher than those of variance of income). Moreover, this can be explained by the importance of

informal sector in the Ghanaian labor market. It is likely that excess supply from the formal sector will seek work in the informal sector. Therefore, migrants know that even if unemployment is high, it is very likely that they will be able to find employment in the informal sector. Finally, we should also notice that migrants arriving into a labor market not only shift labor supply but, if they in fact become entrepreneurs, shift labor demand up.

TABLEAU 2.15 Ratios of Mean Income, Variance in Income andUnemployment

	Full population		
	(1)	(2)	(3)
	Full sample	Men	Women
Alternative-specific variables			
Ratio of Mean Inc.	-19.088***	-20.503***	-18.251***
	(3.488)	(3.939)	(6.129)
Ratio of Var. in Inc.	-23.799***	-21.318***	-26.661***
	(4.214)	(5.142)	(5.442)
Ratio of Unemp. Rate	-2.036***	-1.804***	-2.304*
	(0.508)	(0.391)	(1.325)
Individual-specific variables	Yes	Yes	Yes
Observations	18780	11380	7400
Individuals	1878	1138	740

Note : Robust standard errors are in parentheses and there are cluster at SHS level. *** p<0.01, ** p<0.05, * p<0.1. Anne Duplantier (2021)

2.5.4 Robustness Check

In order to check the sensitivity of the findings, we replicate the analysis adding regional characteristics as independent variables. The community module of the GLSS6 survey includes demographic characteristics of communities, economy and infrastructure, education, health and agriculture. It was administered to only rural communities in the enumeration areas selected for the survey. Questions were asked to chiefs of the communities to collect information on the existence and availability of facilities such as schools, roads, healthcare and utilities such as electricity and water in the communities. As in our 'Go Transition' survey, where we conducted interviews in both rural and urban SHS, we know that this community data has limitations. If most migrants in the 'Go Transition' survey move to cities and not rural communities, these regional characteristics are an imperfect measure of destination characteristics.

The variable *motorroad* is a dummy taking the value 1 if a motorable road goes through the community. The variable *electricity* equals to 1 if the community has electricity or uses a generator or solar energy and *pipeborne* has a value of 1 if any households in the community have access to pipe-borne water or borehole. The variable *mobilephone* is equal to 1 if there is a mobile phone

network in the community. *Permmarket* takes the value 1 if there is respectively a permanent daily market and a periodic market in the community. The variable *pubtransp* identifies if public transport passes by the community. The variables *primschool*, *jhs* and *shstechvoc* are equal to 1 if the community has respectively a primary school, a JHS and a SHS.

TABLEAU 2.16 Utilities and Amenities Characteristics inCommunities (%)

	Ash	Br Ah	Cen	East	Gr Ac	North	Up Ea	Up We	Vol	West	Gha
Motorroad	94.83	85.96	89.06	88.24	90.91	66.32	73.56	89.77	85.00	86.36	83.50
Electricity	75.86	52.63	68.75	47.06	54.55	31.58	52.87	53.41	75.00	59.09	55.86
Pipeborne	24.14	17.54	46.88	18.82	45.45	11.58	21.84	14.94	27.50	12.12	21.45
Mobilephone	82.76	78.95	85.94	67.06	90.91	89.47	86.21	87.50	81.25	51.52	79.74
Permmarket	27.59	14.04	23.44	8.33	9.09	1.05	5.75	35.23	6.25	9.09	13.77
Pubtransp	75.86	70.18	78.13	62.35	72.73	29.47	43.68	50.00	63.29	69.70	58.12
Primschool	84.48	82.46	75.00	60.00	36.36	64.21	73.56	78.41	80.00	83.33	74.10
Jhs	74.14	59.65	53.13	45.88	27.27	31.58	49.43	61.63	66.25	60.61	53.99
Shstechvoc	12.06	1.79	15.63	2.41	9.09	1.05	20.69	24.14	16.25	3.03	11.06
Communities	58	57	64	85	11	95	87	88	80	66	691

Anne Duplantier (2021)

Table 2.16 and Figure 2.7 present the percentage of rural communities, per region and in Ghana, having characteristics described above. The majority of rural communities have a motorable road extending though the community in every region. Only in Upper East and Upper West do fewer than 80% of rural communities have a motorable road. Concerning access to electricity and pipe-borne water, there are more disparities across regions. In the poor Upper East region, only 31.58% and 11.58% of communities have access to these amenities, whereas in the Central region 68.75% and 46.88% of communities having electricity and pipe-borne water.

Table 2.17 shows that the sign and significance of interest variables (the ratio of expected income and the ratio of variance in incomes) stay stable and are not affected by the inclusion of regional characteristics. Even if some regional characteristics of rural communities have a significant effect on the migration decision, the expected mean income and variance in incomes are still attracting migrants. When the relative access to electricity and the mobile phone network increase in the origin region, the probability to leave this region decreases. Similarly, an increase of relative numbers of primary school and SHS in the destination region increases the probability to migrate to this region. The results are similar even when we disaggregate by gender.

As a last robustness check, we decide to take into account the average level of poverty in each region. Indeed, it is possible that individuals have a general idea about the poverty level of the different regions in Ghana. While taking the decision to migrate in a region, they probably keep in mind this information. Through the GLSS6 database, we have access to the poverty lines computed by the Ghana Statistical Office. Therefore we are able to capture the poverty level of each region and integrate it into our model.

FIGURE 2.7 Community Characteristics

TABLEAU 2.17 Ratio of Expected Income and Ratio of Variance in Income With Regional Characteristics - GLSS6

	Full population			Youth with SHS		
	(1)	(2)	(3)	(4)	(5)	(6)
	Full sample	Men	Women	Full sample	Men	Women
Alternative-specific variables						
Ratio of Exp. Inc.	-15.60***	-18.96***	-10.42**	-5.38***	-3.40***	-9.35***
	(3.13)	(3.63)	(5.33)	(0.65)	(1.21)	(2.23)
Ratio of Var. in Inc.	-16.96***	-14.01***	-19.12***	0.002	-0.48*	0.13**
	(3.20)	(3.22)	(4.72)	(0.29)	(0.288)	(0.05)
Ratio of Motorroad	66.74***	59.48*	113.97***	20.39	21.90	16.92
	(25.73)	(35.75)	(52.49)	(22.05)	(35.66)	(40.25)
Ratio of Electricity	11.79***	-12.50***	-12.24***	-4.85**	-8.21**	-10.03***
	(2.25)	(2.72)	(4.01)	(2.21)	(3.22)	(3.52)
Ratio of Pipeborne	-0.29	0.21	-1.19	-0.15	0.65	-2.98***
	(0.53)	(0.58)	(1.21)	(0.63)	(0.80)	(1.15)
Ratio of Mobilephone	-12.81***	-13.25***	-10.75***	-16.91***	-17.11***	-11.80***
	(2.19)	(2.38)	(3.26)	(2.35)	(2.64)	(3.66)
Ratio of Permmarket	-0.61	-0.99**	-0.44	-0.54	-0.74	-0.31
	(0.43)	(0.48)	(0.35)	(0.43)	(2.425)	(0.37)
Ratio of Pubtransp	0.02	5.74	-14.67	-2.15***	4.53	-18.68***
	(4.68)	(4.43)	(9.04)	(5.19)	(3.56)	(6.63)
Ratio of Primschool	-7.84***	-8.52***	-6.35***	-6.08***	-4.82**	-7.50***
	(1.87)	(2.51)	(2.39)	(1.75)	(2.27)	(2.76)
Ratio of Jhs	2.41*	2.71*	2.15	0.31	0.39	2.39
	(1.32)	(1.61)	(1.85)	(1.22)	(1.67)	(2.06)
Ratio of Shstechvoc	-0.17***	-0.11*	-0.43***	-0.18***	-0.12	-0.20
	(0.04)	(0.06)	(0.09)	(0.06)	(0.10)	(0.14)
Individual-specific variables	Yes	Yes	Yes	Yes	Yes	Yes
Observations	18470	11210	7260	18470	11210	7260
Individuals	1847	1121	726	1847	1121	726

Note : Robust standard errors are in parentheses and there are cluster at SHS level. *** p<0.01, ** p<0.05, * p<0.1. Anne Duplantier (2021)

We compute the percentage of poor households in Ghana and by region, given two types of poverty line : the absolute poverty line and the extreme poverty line. The absolute poverty line indicates the minimum living standard in Ghana by including both food and non-food consumption. According to the Ghana Statistical Service (2014b), the absolute poverty line is about 1314 GHS per equivalent adult per year in the January 2013 prices of Greater Accra Region, which corresponds to 1.83 USD per day. In the whole country, 22.5% of the population has a consumption level below this absolute poverty line.

The extreme poverty line reports the consumption expenditure for a minimum food basket providing 2900 calories per adult equivalent per day. A level of consumption below the extreme poverty line indicates that even if a household spends their entire budget on food, they still do not meet the minimum calorie requirement. According to Ghana Statistical Service (2014b), the extreme poverty line of 792.05 GHS per equivalent adult per year in the January 2013 prices of Greater Accra Region, i.e. 1.10 USD per day. In the whole country, 8.9% of the population has a consumption level below this extreme poverty line.

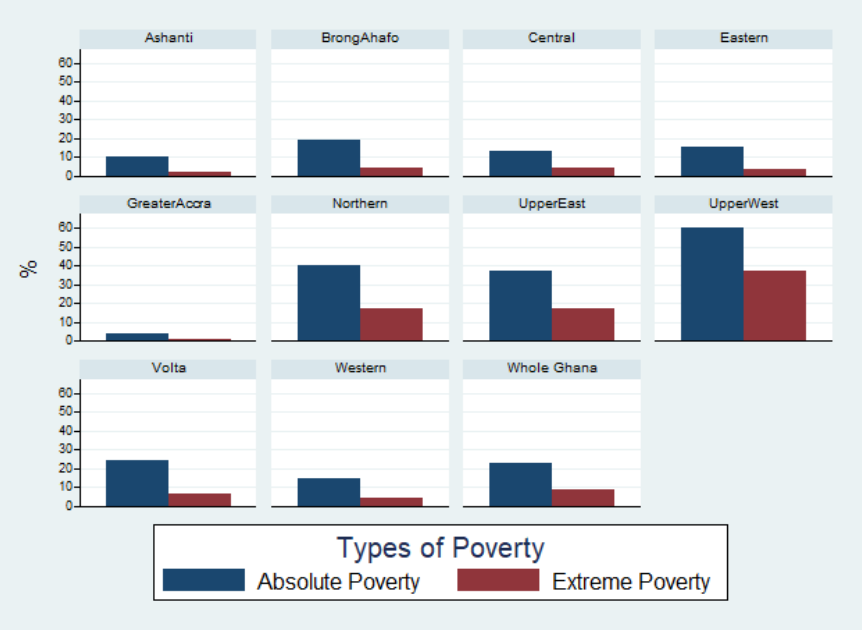
TABLEAU 2.18 Ghanaian Household Poverty (%)

	Absolute Poverty	Extreme Poverty
Ashanti	10.40	2.03
Brong Ahafo	19.14	4.61
Central	13.37	4.27
Eastern	15.39	3.89
Greater Accra	3.59	0.90
Northern	40.48	17.04
Upper East	37.17	17.14
Upper West	60.46	37.05
Volta	24.35	6.40
Western	14.90	4.11
Whole Country	22.59	8.94

Note : Percentage of households (weighted) below the national poverty line, given two types of poverty line : the absolute poverty line in the first column and the extreme poverty line in the second column. Anne Duplantier (2021)

Table 2.18 and Figure 2.8 point out the huge disparities between the regions in terms of poverty. In Greater Accra region, very few households (3.5%) are considered as poor according to the absolute poverty line. In the Ashanti, Central, Western and Eastern regions, the poor households represent around 15% of the population. The regions with the highest rates of households below the absolute poverty line are Upper West (60.4%), Northern (40.4%) and Upper East (37.17%). If we consider the extreme poverty line, the ranking is quite the same. Only 0.9% of the households living in

FIGURE 2.8 Poverty Characteristics



Greater Accra are extreme poor. The extreme poverty rate is between 2% and 5% in Ashanti, Eastern, Western, Central and Brong Ahafo. Three regions (Upper West, Upper East and Northern) stand out with the highest proportion of households in extreme poverty (37.1%, 17% and 17.1% respectively).

Tables 2.19 and 2.20 report the results of the model (for the full population and the sub-sample of youth with SHS) after controlling for the ratios of absolute and extreme poverty, respectively. The sign and significance of interest variables (ratio of expected income and ratio of variance in income) stay stable and are not affected by the inclusion of both poverty variables. Moreover, the ratio of absolute poverty has a positive and statistically significant link on the decision to migrate. When the absolute poverty level increases in the origin region, the probability of leaving this region increases. When looking the sub-sample of youth with education (Columns (4)-(6) of Table 2.19), the ratio of absolute poverty is no longer significant, but the ratio of expected income is still significant and consistent. Concerning the variance, there is a significant link for men only. The results with extreme poverty (Table 2.20) are really close to the ones with absolute poverty. We conclude that our results are robust to the inclusion of poverty level by region.

We have applied two strategies in order to test the robustness of the results. First, we controlled for the regional characteristics of infrastructure, for example the access to a motorable road, or the presence of schools or market in the community. We then took into account the regional poverty level. In the light of this sensitivity analysis, we can conclude that the results are robust and stable.

TABLEAU 2.19 Ratio of Expected Income With Regional Characteristics and Absolute Poverty - GLSS6

	Full population			Youth with SHS		
	(1)	(2)	(3)	(4)	(5)	(6)
	Full sample	Men	Women	Full sample	Men	Women
Alternative-specific variables						
Ratio of Exp. Inc.	-25.051***	-25.875***	-25.507***	-7.600***	-5.162***	-14.445***
	(3.487)	(3.849)	(6.357)	(0.840)	(0.986)	(2.392)
Ratio of Var. in Inc.	-24.101***	-22.773***	-25.497***	-0.137	-1.039***	0.060
	(3.903)	(4.634)	(5.691)	(0.102)	(0.269)	(0.052)
Ratio of Abs. Pov.	0.568**	0.556*	0.772**	-0.216	0.494	-0.471
	(0.240)	(0.297)	(0.337)	(0.323)	(0.343)	(0.547)
Individual-specific variables	Yes	Yes	Yes	Yes	Yes	Yes
Observations	18470	11210	7260	18470	11210	7260
Individuals	1847	1121	726	1847	1121	726

Note : Robust standard errors are in parentheses and there are cluster at SHS level. *** p<0.01, ** p<0.05, * p<0.1. Anne Duplantier (2021)

TABLEAU 2.20 Ratio of Expected Income With Regional Characteristics and Extreme Poverty - GLSS6

	Full population			Youth with SHS		
	(1)	(2)	(3)	(4)	(5)	(6)
	Full sample	Men	Women	Full sample	Men	Women
Alternative-specific variables						
Ratio of Exp. Inc.	-22.834***	-23.816***	-22.183***	-7.438***	-4.961***	-14.389***
	(3.131)	(3.450)	(5.627)	(0.785)	(0.928)	(2.344)
Ratio of Var. in Inc.	-24.287***	-22.965***	-25.940***	-0.148	-1.017***	0.056
	(3.971)	(4.670)	(5.830)	(0.104)	(0.250)	(0.052)
Ratio of Ext. Pov.	0.098***	0.096***	0.122***	0.065**	0.118***	0.035
	(0.022)	(0.028)	(0.032)	(0.031)	(0.034)	(0.060)
Individual-specific variables	Yes	Yes	Yes	Yes	Yes	Yes
Observations	18470	11210	7260	18470	11210	7260
Individuals	1847	1121	726	1847	1121	726

Note : Robust standard errors are in parentheses and there are cluster at SHS level. *** p<0.01, ** p<0.05, * p<0.1. Anne Duplantier (2021)

2.6 CONCLUSION

This chapter studies the internal migration choices of recent SHS graduates in Ghana. It contributes to the literature in development economics by focusing on internal migration choices of educated youth in Ghana, an important topic which has been little studied in the literature. We aim to understand how both expected income and income variability differentials between the regions of origin and destination affect the probability of migrating to any of the regions in Ghana.

A choice model estimated with a mixed logit is used in order to take into account the characteristics of both the alternatives (here regions) and the individuals. Moreover, this model has the advantage to relax the assumption of the independence of irrelevant alternatives of the basic MNLs. This allows for the existence of alternatives with similarities and represents more accurately the real characteristics of the regions in Ghana.

We find that expected income and variance in income have a significant impact on the probability of migrating. However, in the decision to migrate, individuals take the income differentials into account more than the variance differentials. Moreover, there is no significant difference in the decision of the girls and boys. These results are robust when we control for the utilities and amenities of the regions, as well as for the level of poverty in each region. This implies that the educated youth migrate inside Ghana to search for higher average incomes. The results suggest also that they are not afraid of the variability of the incomes.

The understanding of domestic population movement is crucial for allocating resources and for designing policy tools to provide support and incentives or disincentives to migration, if desired. Finally, it is particularly relevant to recognize the returns to secondary education in Ghana and the opportunities open to young high school graduates. This may help the government of Ghana in understanding the effectiveness and relevance of its education system.

CHAPITRE 3
THE EXPECTATIONS AND THE POST-SECONDARY EDUCATION OF GHANAIAN YOUTH

Résumé

Contrairement aux hypothèses de la théorie économique traditionnelle, les informations sur lesquelles les individus prennent des choix sont souvent imparfaites et les attentes jouent un rôle non négligeable dans les processus de décision. Cependant, à cause d'un manque de données d'enquête abordant ces questions, peu d'études empiriques existent dans les pays en développement sur le rôle des attentes dans les décisions d'éducation. Ce chapitre se penche sur le lien entre les rendements espérés de l'éducation et les décisions de postuler et de fréquenter l'université. La méthodologie utilisée est un modèle linéaire de choix. Les résultats indiquent que les rendements espérés de l'éducation sont positivement liés à la décision de postuler à l'université. Cependant, une fois que l'on prend en compte les capacités réelles et estimées, les rendements espérés ne sont plus significativement liés aux décisions d'éducation post-secondaire. Les capacités réelles et perçues sont positivement liées à la probabilité de candidater à l'université ainsi que la décision de fréquenter l'université.

3.1 INTRODUCTION

Contrary to assumptions of traditional economic theory, the information on which individuals base their decisions may often be imperfect. The impact of imperfect information on an individual's decision has been studied by several papers in different areas. These include the role of information [1] on retirement plan decisions in the USA (Duflo and Saez, 2003), as well as on the sexual behavior of teenagers vis-à-vis human immunodeficiency viruses (HIV) risk in Kenya (Dupas, 2011). Álvarez and Vera-Hernández (2013) examine the impact of knowledge [2] on the use of medical care by mothers in Colombia, while McKenzie et al. (2013) analyse the link between information and international migration decisions from Tonga to New Zealand. [3]

With regards to education decisions, students usually make their schooling decisions under limited or imperfect information. On one hand, returns to education are underestimated by the parents and students in Madagascar (Nguyen, 2008) and in Dominican Republic (Jensen, 2010). On an other hand, students from public high schools in Boston overestimate the returns to college (Avery and Kane, 2004). [4]

The impact of information on education choices has been studied recently, with authors finding that, in general, providing information influences education choices. In the USA, Hastings and Weinstein (2008) provide information on school test scores to parents. They find that the information influences parents to choose better-ranked school for their children.

In Dominican Republic, Jensen (2010) provides information to 8th grade students on the returns to education and assesses the long-term impact of this intervention. Bringing information about real returns increased perceived returns of individuals who were re-interviewed four to six months later. Relative to those whitout, students with information completed on average 0.20 more years of schooling over the next four years. In Madagascar, Nguyen (2008) provides parents with information about returns to schooling. The results suggest that after providing statistics, parents update their beliefs about the return of an average child and about the return of their own child. A few months later, children increase their primary school attendance and academic performance.

As for perfect information, economists usually suppose that individuals making a decision have rational expectations (Hurd, 2009). If so, individuals' subjective probability distributions should be the same as the true probability distributions. But as Delavande et al. (2011) explain, the majority of economic decisions are subject to uncertainty and imperfect information. Beyond the general

1. Using field randomized experiment.
2. Using survey data.
3. Using both natural experiment and survey data.
4. Avery and Kane (2004) find that students in Boston public schools tend to overestimate the returns to college, while they are overly pessimistic about costs of college and their own qualification.

benefits and costs, risk preferences and time preferences are not the only determinant of economic decisions. One must also take into account expectations for the future in order to better understand the behavior of individuals and the possible impact of policies.

Delavande et al. (2011) illustrate this issue with the following example. If decision-makers want to implement policies encouraging the youth to go to school, they have to first understand why they do not go. The reasons could be that expectations on return to education are low, or that the cost of education is high. If economists do not capture these expectations, they cannot distinguish the effects of each reason, and thus cannot design effective policies for decision-makers.

In most of the papers on education choices, expected returns to education are proxied by observed earnings (Attanasio and Kaufmann, 2014). This is the method we chose in Chapter 3 to analyze the link between migration decision and income differences between region. Indeed, our data do not include the respondents' expected incomes for each region. But in order to use this method, researchers have to assume that individuals have good knowledge and perfect information about the actual labor market. Another limit of this approach is that researchers only know for each individual the earnings for the specific selected education level and not for the other possible education levels that they could choose.

Expectations data allow us to overcome these limitations by relaxing the previous assumption and obtaining expectations for both chosen and not chosen education levels, called "counterfactual outcomes" by Arcidiacono et al. (2012) and Kaufmann (2014). In the USA, an increasing body of literature addresses expectations using surveys (Manski, 2004). However, few surveys include expectations questions in developing countries (Delavande et al., 2011).

We define expectations as the subjective perception of something. In this chapter, we consider several types of expectations. Firstly, the expected earnings are captured through the perceived earnings distribution they will get in 10 years in a given situation. We also consider the expected academic results of the respondent by asking them the results they think they will get at their final exam. Then, we capture the expected ability of the respondent to get into university : do they think their scores will allow them to enter university ? Finally measure their perceived financial ability to afford university. On a scale from 0 to 10, they have to estimate how likely they think the event will happen.

This chapter examines the role of expectations on schooling decisions in Ghana. Do the expected returns to schooling explain the post-secondary education choices ? What is the role of the student's ability in their decision to apply to or attend university ? To answer these questions, the data collected and presented in Chapter 1 are used. In particular, I have information on the expectations of respondents when they were still at senior high school (SHS).

The chapter looks at whether students' expectations are correlated to the decision to continue to university once they have graduated from SHS. One contribution is to decompose the post-secondary education process in two steps : application to university and attendance at university. This study contributes also to the literature by including academic and perceived ability to the estimation of the choice of education. The results show that the expected returns to university are positively associated to the decision to apply to university. The evidence is less strong for the attendance to university. Moreover, after controlling for ability, the relationship is not significant anymore.

The chapter is structured as follows. First, a review of the literature on expectations and schooling is presented in Section 2. Then Section 3 sets up the theoretical framework of the chapter, and Section 4 presents the data used for this study, as well as the way in which the variables were built. Section 5 sketches the descriptive statistics about the expectations and the education characteristics of the sample. The empirical methodology is discussed in Section 6, while the results and conclusion following in Section 7.

3.2 LITERATURE

This section reviews the economic literature on expectations. First I present how expectations are used in the economic literature. Then, the literature studying the link between expectations and schooling is reviewed. The third part of this section sheds light on the way the expectations are measured in the economic literature. Finally, I explain how the chapter is related to this literature.

3.2.1 The Collection of Expectations in Economics

The use and the measure of expectations in economics has been questioned for a long time. Several works, as such Manski (2004), Attanasio (2009), Hurd (2009), Delavande et al. (2011) and Zafar (2011a), focus only on the question of the measure of expectations. These authors distinguish two types of measure of expectations : the non-probabilistic methods using a Likert scale [5] and 'what do you think/expect' questions, and the probabilistic methods. The former were previously used more than statistical concepts, as they seem to be easier to understand for respondent. Researchers also fear that quantitative measures of probabilities take too much time in the interview process or that the answer will have little value. But one limit of this commonly used method is that the answers of these non-probabilistic questions might be interpreted differently according to the respondents. The consequence is the difficulty in making inter-person comparisons.

One major concern about using a probabilistic measure of expectations is that the respondent may not understand the concept of probabilities. To be sure that the respondent understands well the

5. The Likert scale uses graduated answers such as 'very likely', 'likely', 'not very likely'.

concept of probability, Attanasio (2009) recommends using examples as the probability of rain the next day, as well as visual aids asking the respondent to allocate ten stones, balls or beans that represent probabilities. In rural Malawi, Delavande and Kohler (2009) find that respondents allocating beans seem to understand very well the concept of probability. Attanasio (2009) finds the same results in Colombia. In a review of the literature on measuring expectations, Delavande et al. (2011) conclude that expectations are well understood by respondents and do not take much more time. They strongly encourage researchers to use expectation questions in developing country surveys. Following this literature, we chose to apply the probabilistic method to collect expectations.

Attanasio (2009) recommends some procedures to collect subjective expectations data. For discrete variables, it is common to ask the probability of a given realization. For continuous variables, it is more challenging as one searches the probability distribution. A solution is to elicit the range of variation of the variable, in asking the minimum and maximum values at a future date. Combined with the probability assessed to certain sub-intervals, [6] the minimum and maximum give some points of the cumulative distribution function of the variable. To use these, an assumption about the form of the distribution is necessary, for example log-normality, piecewise linearity and triangular as seen in the literature. The survey that I used in this chapter follows some of these procedures. I explain the method in more detail in Section 4.

As Delavande et al. (2011) explain, it is difficult to know if expectations collected by survey are reliable, as what people think is not directly verifiable. Attanasio (2009) suggests to check the consistency of these data with both internal and external validation. For internal validity, one method is to check if the expectations vary in the same direction as observable characteristics. For example, they propose to check how the mean of future incomes covary with the education achievement of respondents. [7] For external validity, one can test if expectations and actual realizations are systematically different, as long as having these data for several periods. But it is quite difficult to obtain such data and even if one does, there are several reasons why expected incomes differ from realized incomes, such as selection into education of individuals with different degrees of ability.

3.2.2 Expectations and Schooling

The literature on expectations and schooling started in the 1990s by focusing on the USA. Smith and Powell (1990) analyze how students from mid-western state universities form their income expectations, focusing on, among other things, peers' expectations. They find that the type of school

6. Respondents are questioned about their probability to earn more than a fixed value between the minimum and maximum values.

7. Attanasio (2009) assumes that more educated individuals have higher expected incomes.

attended and the father's income influence students' expected incomes. Betts (1996) finds that education level and an individual's characteristics are correlated to University of California (San Diego) students' knowledge of average incomes at national level (and not their own incomes). Dominitz and Manski (1996) is the first study asking high-schools seniors and college freshmen (in Madison, Wisconsin) to elicit their beliefs about the distribution of earnings for different education levels. However, they did not investigate their link to schooling decisions. After presenting the design of the survey, the measure of the expectations and the data collection, they describe the beliefs of their sample of 110 American students about earnings distributions and the returns to college.

Following Manski (2004), more recent literature highlights that expectations are useful for analyzing education choices. Several studies, such as Zafar (2011b), Arcidiacono et al. (2012), Stinebrickner and Stinebrickner (2013) and Wiswall and Zafar (2014), focus on the decision of a college major and find that expected incomes have a significant impact on this choice, but less than other determinants such as beliefs about one's own ability.

Another part of the literature focuses on the link between expected returns to education and schooling decisions. The first two articles using expected returns to education focus on the USA (Rouse, 2004; Avery and Kane, 2004). Rouse (2004) explores the link between expectations and educational attainment of low-income ethnic-minority students from Baltimore and high-income white students from Madison. Rouse (2004) finds that expected returns to post-secondary education are quite similar between the two groups of students. She also analyzes the determinants of both the probability to plan to attend college and of college attendance. The findings suggest that a rise in expected college returns increases the probability to plan to attend college and to actually enroll in college. However, no significant differences are found according to the race and family income of the students.

Several other studies seek to determine whether or not expectations have an impact on education choices in developing countries. Kaufmann and Attanasio (together and apart) wrote several papers about expected returns to education in Mexico using the same survey data.[8] At first, they analyze how students' or parents' expectations impact the decision to attend both high school and college (Attanasio and Kaufmann, 2009). They deepen the analysis in identifying two types of expectations : expected returns and risk perceptions with regard to the labor market.[9] The findings reveal that risk perceptions of earnings are important determinants of high school attendance. Moreover, expected returns matter for decision to attend college. Finally, the authors show that credit constraints also have a substantial impact on college enrollment.

8. This survey data has been collected during the evaluation of the "Oportunidades" (or PROGRESA) conditional cash transfer program.

9. Specifically, earnings and unemployment.

In a closely related work following the same approach, Attanasio and Kaufmann (2014) study the link between expectations and education decision, focusing on the differences between girls and boys. Keeping the same expectation types, [10] they analyze two enrollment decisions : in high school and in college. Results suggest that labor market expectations, both expected monetary returns and risk perceptions, matter in the decision to attend SHS and college. However, schooling decision differs according to the gender of the student. Expected returns to college significantly determine boys' decisions to enroll in college but have no impact on girls' decisions. In contrast, mother's expectations influence decisions to enroll in college among girls but not boys.

With the same data from Mexico, Kaufmann (2014) tries to explain why poor and rich students have very different behavior in college enrollment. The author finds that poor students need higher expected returns to choose to attend college than rich ones. In particular, Kaufmann (2014) also analyzes credit constraints and time preferences. The results suggest that cost constraints matter in the differences between college enrollment of poor and rich students, but not time preferences.

Analyzing the formation of expectations in India, Sequeira et al. (2016) explore two mechanisms : recognition for schooling performances (direct effect) and exposure to successful students (peer effects [11]). The authors use a regression discontinuity design [12] of two fellowship programs that award academic performances in secondary schools. They find that students recognized for their schooling performances have higher expectations on returns to high education. However, the results do not show significant peer effects : students whose friends, neighbors or siblings are awarded do not change their expectations.

Finally, Attanasio and Kaufmann (2017) use the same framework [13] as Attanasio and Kaufmann (2009), Attanasio and Kaufmann (2014) and Kaufmann (2014) in order to bring another view on this research question. They explore the role of expected marriage market returns to explain the college enrollment decisions (in addition to the expected labor market returns). They find that both types of expectation are significant determinants of college attendance, but boys give more weight to labor market returns (versus marriage market returns) than girls.

3.2.3 The Measure of Expectations

This part describes the way in which the questions are asked to elicit expectations in the literature on expected returns to education. In general, respondents are asked to estimate their earnings or others' at age 25 or 30-40, for different levels of education (primary, secondary and university). For

10. Expected returns to education and perceived risks of both unemployment and earnings.
11. The authors wonder whether having a relative awarded with a fellowship has an impact on expectations and education choices of the individuals.
12. The cutoff is the score allowing the award.
13. Same data and same measures of expectations.

this purpose, Avery and Kane (2004), Nguyen (2008) and Jensen (2010) [14] use questions beginning with “How much do you think you will earn ?". These questions have several limitations. First, they are not very accurate and do not elicit the distribution of expected earnings, and so researchers cannot know whether the respondent refers to the mean, median or mode. Second, the question does not take into account inflation, [15] nor uncertainty, nor the lifetime profile of earnings. Third, the questions can be too long and complicated for young students. [16]

In order to overcome some limitations pointed out above, researchers try to be more accurate. Concerning the precision, Rouse (2004) elicits the expectations of the median by asking for the best guess as to the amount of money the respondent thinks they will earn per year at 30/40 years. Rouse (2004) also defines the concepts of median and ‘best guess’ to increase chances that students understand the statistical concepts. Concerning the issue of lifetime profile, Schweri and Hartog (2017) capture a few points of the expected lifetime earnings in order to approximate under some assumptions the expected lifetime rate of returns. To minimize errors, Betts (1996) and Sequeira et al. (2016) choose to present to respondents a list of values rather than letting respondents freely estimate the earnings. The understanding of respondents is less questionable when they are students in post-secondary education, as in Schweri and Hartog (2017).

Several researchers tempt to capture the distribution of expected earnings. Dominitz and Manski (1996) adopt a hybrid approach, asking first for the median earnings and then for the probabilities that earnings would exceed a sequence of thresholds. Sequeira et al. (2016) capture average expected earnings but also other moments of the distributions. Attanasio and Kaufmann (2009), Attanasio and Kaufmann (2014), Kaufmann (2014) and Attanasio and Kaufmann (2017) use the same questions to obtain the distribution of expected earnings. They first ask questions about both the minimum and maximum monthly income the respondent would earn at age of 25, conditional on a given education level (JHS, SHS and college). They then ask the respondent to estimate the probability that they earns more than the midpoint for this education level. Attanasio and Kaufmann [17] also capture the perceived risk of unemployment by asking the probability of working at age of 25 for the same education levels. This chapter follows the methodology of Attanasio and Kaufmann and captures under certain assumptions the distribution of expected earnings.

In this part, I will review the different ways to compute and use expectations in the literature. The easiest way is to use directly the expected earnings elicited in the questions without transformation. This approach has been chosen by Smith and Powell (1990), Betts (1996) and Avery and Kane

14. Only Nguyen (2008) and Jensen (2010) elicit expected earnings for primary education as their respondents are at primary school.

15. Smith and Powell (1990) ask respondents to assume that there are no inflation.

16. Jensen (2010) interviews children in primary school and 10% of students did not respond.

17. Attanasio and Kaufmann (2009), Attanasio and Kaufmann (2014), Kaufmann (2014) and Attanasio and Kaufmann (2017).

(2004). Alternatively, Sequeira et al. (2016) ask respondents to assign subjective probabilities to receiving earnings in different bands. [18] Then, expected income for a given schooling level is computed by weighting each income band (using the lower bound) with its perceived probability. On the other hand, Rouse (2004), Nguyen (2008) and Schweri and Hartog (2017) elicit the expected returns to education by taking the ratio of expected earnings.

Finally, Dominitz and Manski (1996), Jensen (2010), Attanasio and Kaufmann (2009), Attanasio and Kaufmann (2014), Kaufmann (2014) and Attanasio and Kaufmann (2017) choose another method to elicit expected returns. They take the difference between expected earnings of two different levels of education. But in order to use the expected earnings, Attanasio and Kaufmann (2009), Attanasio and Kaufmann (2014), Kaufmann (2014) and Attanasio and Kaufmann (2017) have to compute it first. Indeed, they do not ask directly 'what the expected earnings are', but instead ask for some points of the distribution (minimum and maximum) and the probability to earn more than the midpoint. In order to capture the expected earnings, they make an assumption on the functional form of the distribution (they use three functional forms : step-wise uniform, bi-triangular and triangular). The same method is used in this chapter and is explained latter.

3.2.4 Position in the Literature

I explain here how this work is related to the literature and what are the contributions. In terms of data collection, the survey follows the recommendation of Delavande et al. (2011) and uses the probabilistic method in order to elicit expectations. Following Attanasio (2009), the survey includes a section titled 'Probability Practice' allowing to explain the concept of probability with simple examples such as the likelihood of rain. This section also allows us to check whether or not the respondents understand well the concept of probability and ensures that the data are reliable. Section 4 of this chapter highlights the 'Probability Practice' in more detail and discusses the reliability of the data on expectations.

Regarding the way to capture expectations, the questionnaire follows the recommendations of Attanasio (2009). As in his paper, our survey asks for the minimum and maximum values as well as the probability assessed to a certain value. These questions allow us to obtain, under certain assumptions, the distribution function of the expectations. In order to measure the returns of an education level, I finally choose the method used by Dominitz and Manski (1996), Jensen (2010), Attanasio and Kaufmann (2009), Attanasio and Kaufmann (2014), Kaufmann (2014) and Attanasio and Kaufmann (2017) that takes the difference between expected earnings of two different levels of education.

18. $y = [0 - 5000; 5001 - 10000; 10001 - 15000; 15001 - 20000; > 20000]$

This research is mainly based on Attanasio and Kaufmann (2014) but differs from the literature in several points. First, this paper proposes two dependent variables of education choices, attendance and application, whereas in the literature the education outcomes is mainly school attendance. [19] Using application in addition to attendance allows us to really capture what students' wishes and aspirations. The attendance alone captures the education choice of the students, but also the selection process of the institution they finally attend.

Second, Kaufmann (2014) highlights that, in addition to student perceptions of the economic benefits to education, it would be pertinent to capture their perceptions about their abilities. To the best of my knowledge, Arcidiacono et al. (2012) is the only paper measuring directly perceived ability, asking students to estimate "their competitiveness relative to their peers at Duke in each of the six majors" (cf Arcidiacono et al. (2012), page 10). To proxy for students' perceptions about their own future performance, Kaufmann (2014) uses information about past school performance. Following the same purpose, I use the academic results to capture the ability of the students. The survey also asks respondents to estimate how likely it is that their West African Senior School Certificate Examination (WASSCE) [20] results will be good enough to get into university. This measure of perceived academic abilities brings some interesting and quite rare information about what students think about themselves. This paper therefore contributes to the literature by providing information about both perceived and academic ability and using it in the estimation of education choices.

Finally, it is quite rare to obtain survey data on expectations in Africa. In the literature, only Nguyen (2008) focus on an African country (Madagascar). The rich panel database 'Go Transition' allows us to shed light on issues that few have studied in the Sub-Saharan region, mainly because of a lack of data, for example on education and the labor market. A greater understanding of expected returns to education and schooling decisions in Ghana will enable more effective policies to be proposed in order to reduce unemployment and contribute to the development of the country.

3.3 CONCEPTUAL FRAMEWORK

This section discusses the factors of education decision that characterize the conceptual framework of the analysis.

The assumptions of this chapter are based on Becker's model of human capital [21] (1962), where education is considered as an investment for the maximization of lifetime utility. In this frame-

19. To the best of my knowledge, only Avery and Kane (2004) also focus on application process in a descriptive paper.

20. As mentioned in the previous chapters, the WASSCE is a test at the end of SHS.

21. Another possible theorical foundation is the screening theory of education which suggests that educated individuals have higher earnings because of their diplomas which serve as a signal of higher productivity (Spence, 1973; Arrow, 1973).

work, individuals (or families) weigh the current cost of an additional year of schooling against the perceived future income and other benefits. Following this idea, I assume that youth in Ghana make decisions to pursue schooling after SHS by estimating the current costs and future benefits of these additional years of study. As individuals do not know their future benefits, because they cannot know precisely their future actual earnings, I suppose they estimate it by considering their expectations of the labor market and their perceived abilities.

The education choice can be decomposed in two steps : first, the application to an institution and second, the attendance. An SHS student will apply to and then attend a university if they estimate that they have a higher probability to find a job and have higher earnings than if they did not go to the university : the student will estimate the expected earnings conditional to the probability to be employed. Before applying, they will also consider whether or not they feel able to successfully complete a university program, as well as their perceived probability to be able to be admitted into university with their WASSCE score and their perceived ability to afford their attendance to a university.

However, there are several barriers a student faces in relation to attendance that they do not face in the application process. A first limit is the cost implied by the attendance of a post-secondary institution. The monetary costs include the annual tuition and the cost of living. The living fees (such as housing, transport and food) depend on the location of the post-secondary institution ; these will be higher should a student not be able to live with their parents. All these costs will weigh more on the budget of poorer households. The success of the student can also have an impact on the costs of education. Indeed, the best students will attend public university as a non-fee-paying students while others either attend public university as fee-paying student or have to pay substantial fees for education at a private one. [22] In order to capture the monetary constraint, I control in the model for the education level of the parents (as a proxy of the wealth level of the household). I assume that the more educated the parents are, the less the monetary constraint should be as they will likely hold better paid jobs.

Another cost to consider is the opportunity cost to attend a post-secondary institution. This is the income that the student does not earn while they are studying. Even if the student is working part-time during their studies, their pay will likely be lower than what they could earn by working full-time. Depending on the length of the study before graduating, a choice of post-secondary education will have different impacts in terms of monetary costs and costs of opportunity. In general in Ghana, it takes six years to become a physician or dentist, while it takes three years to obtain a teaching and nursing diploma. The diplomas and certificates take two years and the bachelor's degree is

22. There are 7 public universities in Ghana and 21 private institutions officially accredited by the government. The public universities are known to be better than the private ones.

four years. Both the monetary and opportunity costs will usually be higher for long post-secondary education.

This opportunity cost also depends on the preference for the present and the risk aversion of the respondent, particularly where they are certain to obtain a low-skilled job now, for example where a person knowing that they could work at her parents' farm or shop. An SHS graduate with a strong preference for the present will choose to work sooner in order to reduce their opportunity cost, even if their income could be higher after graduating from a post-secondary institution. As there is an uncertainty about the capacity to find a job after graduating from a post-secondary institution, a risk-averse individual will prefer to get a job now if they have the opportunity, even with a lower income. It is difficult to find a good way to take into account these opportunity costs as the preference for the present and the risk aversion are not elicited in the database.

A second limit to attendance is the requirements and selection from the post-secondary institution, as well as the schooling outcomes of the respondent. Moreover, some of these barriers do not depend on the respondent and are exogenous (such as fees and requirements of the institution) while others are endogenous (such as wealth level and schooling results of the respondent). This implies that the choice to attend one specific post-secondary institution over another could not only be linked to earning expectations and employment probability, but also to these exogenous and endogenous constraints. That is why the WASSCE result is included in the model.

While the application process is less affected by these costs, it nevertheless brings others. The first cost is the time cost. Depending on the number of applications, the different steps needed before to apply (such as searching information about the institution and the program), writing statement of purpose letters, collecting the documents required and finally applying, can be costly in time. Moreover, Avery and Kane (2004) highlight that these steps are actual barriers preventing poorer students from applying and pursuing studies in the USA. These different procedures could be further complicated depending on the level of administrative skills of the student and the support the student receive from her acquaintances. These 'administrative costs', as I will call them, could be proxied by the education level of the parents.

Another constraint during an application is the monetary costs. Thse include the application fees that a student has to pay when applying as well as the additional expenses that can arise during the steps of application. In the database, the average fees for application made by the respondents are 79 Ghana cedis (GHS) per application between 2011 and 2014. This represents around 3% of the GDP per capita, which was about 2400 GHS in 2011 according to Duflo et al. (2017). The final limitation of an application could be self-censorship or family pressure. If an individual thinks that they do not fill the requirements for a post-secondary institution, they will not even try to apply. This mechanism will depend on the information and the perceptions of the respondents about the

institution in question, as well as the perception they have of their abilities, skills and schooling results. Someone who is not confident about their chances of success will not apply to a post-secondary institution, even if their expected earnings for this type of education are high. This is the reason why several proxies for the perceived ability of the respondent are included in the estimation of the application choice.

3.4 DATA

This section presents the sample used in this chapter. The database is thoroughly described in Chapter 1. In the sub-section titled 'Probability Practice', I discuss the validity and reliability of the data on expectations using some descriptive statistics. The last part of the section explains in more detail how the expectations are measured.

3.4.1 The Sample

The sample comes from several waves of the Ghana Opportunities for Transitioning Senior High School Students panel database presented in Chapter 1. In 2011, the 'Go Transition' questionnaire collected expectations of respondents from the 2005, 2006 and 2007 cohorts. The same year, the 2008, 2009 and 2010 cohorts were interviewed at school with the 'In School' questionnaire, that included questions about expectations. However, data from the 2011 'In School' wave will not be used here. Indeed, the respondents in this wave were asked to estimate the earnings and probability of employment in five years' time, and not in ten years' time as in other waves.

In 2012 and 2013, the same questions on expectations as in the 2011 'Go Transition' wave were asked to the 2009, 2010 and 2011 cohorts with the 'In School' questionnaire. I could therefore keep the data from the 2011 'Go Transition' wave in the sample. However, the expectations recorded in the 'Go Transition' wave are not expectations in the true sense, as we ask the question to respondents after they had graduated from SHS, chosen a post-secondary institution or started to work. In order to avoid the effects of ex-post rationalization, whereby a respondent may justify their choice by adjusting their expectation ex-post, it seems better to not use the 'Go Transition wave. I therefore use the 2012 and 2013 'In School' waves only.

The sample comprises three cohorts - 2009, 2010, 2011 - for which both the education choices and expectations questions about the labor market and ability were observed in the database. While the original database is a panel database, I do not use the panel dimension here. The main reason is that the interest variables - education choices - do not vary over time. For each respondent there is then only one observation including the expectations and the education choices. The 2009 and 2010 cohorts were interviewed in both the 2012 and 2013 'In School' waves, whereas the 2011

cohort was only interviewed in 2013. I choose to keep the information from 2012 for the cohorts with two observations. Indeed, the answers from 2012 will be more 'pure' and not impacted by a 'survey effect'. In 2013, one can imagine that the respondents had time to inform themselves after they were asked the question in 2012. Therefore, it is better to not use this wave for the 2009 and 2010 cohorts.

3.4.2 Probability Practice

The questionnaire includes a section of probability practice as recommended by Delavande et al. (2011). The respondents were asked to estimate the probability of events such as weather, the sex of a child or a personal life event such as going to the market or getting married. [23] These questions allows us to check how well the respondents understand the concept behind probabilities.

Table 3.1 presents the responses to the questions of probability practice and Table 3.2 shows the percentages of consistent responses. It shows that the probability that a baby will be a boy is 49.77%, which is very close to 50%, and 74.79% of the respondents stated that this probability exactly equals to 50%. The probability that it will rain the next day is further from 50% as the question was not asked to all participants at the same time of year. The answers stay consistent as the average probability that it will rain on a rainy day is larger than that of it raining the next day (70.27% versus 43.4%). Furthermore, 90.43% of the respondents stated that the probability of rain is higher for a rainy day than for the next day. Finally, the probability to be married in ten years is consistently higher in average than the probability to be married in five years [24] and 92.56% of the respondents respect this ranking. In general, then, the probabilities are rational and seem to be well understood by the respondents.

TABLEAU 3.1 Probability Practice

Variables	Average	Stand. dev.	Observations
Rain tomorrow	43.4	(24.44)	5462
Rain rainy day	70.27	(21.54)	2216
Baby is a boy	49.77	(12.2)	5462
Market tomorrow	34.4	(28.91)	2218
Married in 5 years	49.45	(27.82)	2181
Married in 10 years	70.23	(25.27)	5423

Note : The probabilities are percentages. The questions "Rain rainy day", "Market tomorrow" and "Married in 5 years" were not asked in the 'In School' waves. That is why the number of observations is varying. Anne Duplantier (2021)

23. Detailed questions are in Section B.1 in the Appendix.

24. Divorce rates are very low in Ghana. In 2014, 3% of the people aged 15-49 were divorced (Ghana Statistical Service et al., 2015).

TABLEAU 3.2 Consistent Responses during Probability Practice

	Percentage	Observations
“Rain rainy day" $>$ 50	75	2216
“Baby is a boy" = 50	74.79	5462
“Rain rainyday" $>$ “Rain tomorrow"	90.43	2216
“Married 10yrs" $>$ “Married 5yrs"	92.56	2178

Note : In this table I count the number of consistent responses. Anne Duplantier (2021)

In order to know how well the respondents understand the concept of probability and whether or not the answers to the questions of expectations can be trusted, we also asked questions to surveyors. At the end of the questionnaire,[25] we ask the surveyor : “Did you have to explain the concept of probability more than once ?" and “Did the respondent appear to understand the concept of probability ?". Table 3.3 presents the percentage of respondents by survey and cohort who needed more than one explanation about the concept of probability. The understanding of the respondents is also presented using a scale of four degrees : ‘not at all’, ‘not so well’, ‘well’, ‘very well’.

In 2012, about one out of two respondents asks the surveyor to repeat the explanation of the probability concept (46% for the older cohort (2009) and 53% for younger (2010)). In 2013, this proportion decreases considerably : only 28% of the older cohorts need more than one explanation, 42% of the 2010 cohort and 37% of the 2011 cohort. This could be explained by the fact that the respondents of the 2009 and 2010 cohorts were familiar with the probability as they were interviewed the previous year. Another explanation is that the respondents are simply older than in 2012 and understand the concept better.

The surveyors’ responses to the second question show that a very large majority of respondents appear to understand well or very well the concept of probability. In 2012 between 86% and 88% of the respondents understood well or very well the concept, increasing to 93% for the three cohorts in 2013. These results suggest that the surveyors estimate that the respondents seem to have a good understanding about the concept of probability. This confirms the results of the probability practice questions. In conclusion, I can use the expectation variables without a major concern about the understanding of the respondents.

3.4.3 Capturing the Expectations

The Questions on Expectations

The questions on expectations on labor market outcomes are based on the work and the questions

25. In the 2012 and 2013 ‘In School waves.

TABLEAU 3.3 Understanding Evaluation

Survey	In School 2012		In School 2013		
Cohort	2009	2010	2009	2010	2011
Need to repeat	46.26	53.55	28.06	42.64	37.52
Resp. understands					
Not at all	0.80	0.14	0.00	0.00	0.00
Not so well	10.96	13.11	6.78	6.50	6.05
Well	43.85	46.04	35.59	39.01	40.00
Very well	44.39	40.71	57.63	54.49	53.95
Observations	748	732	531	523	645

Note : Percentage of respondents. Anne Duplantier (2021)

of Attanasio and Kaufmann (2009, 2014, 2017) and Kaufmann (2014). In order to capture the perceived risk of unemployment, respondents were asked to estimate the probability of working in ten years as wage-employed or self-employed, for each type of post-secondary institution (university, polytechnic, teacher training, nursing or no post-secondary education). The question is : “Out of ten, if you attend [University/...], how likely is it that - in ten years’ time - you will be working, either wage-employed or self-employed ?" As set out by Attanasio and Kaufmann (2014), I capture the earning expectations by following several steps. First, respondents estimate both the minimum and the maximum amount earned in ten years if they are waged- or self-employed, depending on the type of post-secondary institution attended. The questions are : “Think about if you attend [University/...] and - in ten years’ time - are wage-employed or are self-employed, what is the minimum amount you would earn per month ?" and “What is the maximum amount you would earn ?"

Then, the midpoint of the maximum and minimum amount is computed. Finally, the interviewer asks the respondent to estimate the probability of earning more than the amount computed as the midpoint (assuming the same conditions defined previously). The question is : “If you attend [University/...] and - in ten years’ time - are wage-employed or are self-employed, out of ten, how likely is it that you will make more than [amount equals to the midpoint] ?" In the questionnaire, the questions are asked differently for the respondents who have already attended or are attending a post-secondary institution during the interview. For the ‘No post-secondary education’ category, the survey asks the respondent for their employment probability and earnings expectations when they do not attend a post-secondary institution after SHS (“If you do not study after SHS").

In addition, the survey includes questions on other types of expectations. First, the respondents are asked to estimate the cost of attendance, i.e. tuition fees, for each level of post-secondary education with the question : “How much would the program cost per year if you went to a public university as a regular student ?" I then capture the perceived ability of respondents using two different

questions. Respondents are asked to estimate their expected results to WASSCE before they take the exam : “What aggregate score do you expect to achieve on the WASSCE next year ?" Another question measures their perceived ability to be admitted into university with their WASSCE results : “Out of ten, how likely is it that your WASSCE results will be good enough for you to get into university without having to take remedial WASSCE classes ?" In the last question, respondents have to estimate their financial ability to afford university : “Out of ten, how likely is it that you would be able to afford to attend university as a regular (non-fee paying) student if you were admitted ?"

The Measure of Expectations

Following Attanasio and Kaufmann (2014), I seek to capture the subjective distribution of future earnings $f(y)$ for an individual who attended university. The survey provides for each respondent information on the support of the distribution $[y_{min}, y_{max}]$, the midpoint $y_{mid} = (y_{min} + y_{max})/2$ and the probability to earn more than the midpoint $p = Pr(y > y_{mid})$. An assumption on the distribution function $f(.)$ is needed. Attanasio and Kaufmann (2014) use three different functional forms for $f(.)$: step-wise uniform, bi-triangular and triangular. In the data, the estimated average probability (p) to earn more than the midpoint (y_{mid}) is superior to 50%. Therefore, the step-wise uniform distribution and the U-shape function are excluded and the triangular distribution is chosen (See Section B.2 in the Appendix for a detailed explanation of the function). In this functional form, there is more weight to the middle of the distribution and less to the extremes, as shown in Figure 3.1.

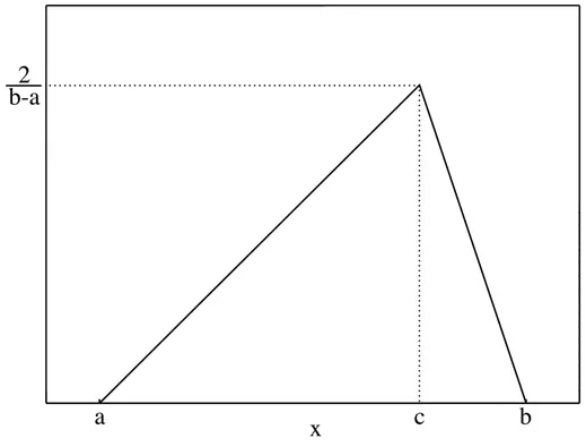
Let y be a continuous random variable varying between y_{min} and y_{max}. If $f(y)$ is a probability density function, the cumulative density function is therefore $F(y) = \int_{y_{min}}^{y_{max}} f(y).dy$. One can interpret density functions as probabilities. If the density function $f(y)$ follows a triangular distribution, then the probability that y be superior to y_{mid} is :

$$p(y > y_{mid}) = p(y_{mid} < y < y_{max}) = \int_{y_{mid}}^{y_{max}} f(y).dy \tag{3.1}$$

Therefore, the probability density function is :

$$\begin{aligned} f(y|y_{min}, y_{mid}, y_{max}) = & \quad 0 \qquad\qquad \text{if } \; y < y_{min} \\ = & \quad \frac{2(y - y_{min})}{(y_{max} - y_{min})(y_{mid} - y_{min})} \quad \text{if } \; y_{min} \leq y \leq y_{mid} \\ = & \quad \frac{2(y_{max} - y)}{(y_{max} - y_{min})(y_{max} - y_{mid})} \quad \text{if } \; y_{mid} \leq y \leq y_{max} \\ = & \quad 0 \qquad\qquad \text{if } \; y > y_{max} \end{aligned} \tag{3.2}$$

FIGURE 3.1 Triangular Probability Density Function



Source : Mathworks

And the cumulative density function is :

$$F(y|y_{min}, y_{mid}, y_{max}) = \int_{y_{min}}^{y_{max}} f(y).dy \tag{3.3}$$

That is :

$$\begin{aligned} F(y|y_{min}, y_{mid}, y_{max}) = & \quad 0 \quad \text{if } y \leq y_{min} \\ = & \frac{(y - y_{min})^2}{(y_{max} - y_{min})(y_{mid} - y_{min})} \quad \text{if } y_{min} < y \leq y_{mid} \\ = & 1 - \frac{(y_{max} - y)^2}{(y_{max} - y_{min})(y_{max} - y_{mid})} \quad \text{if } y_{mid} < y < y_{max} \\ = & \quad 1 \quad \text{if } y \geq y_{max} \end{aligned} \tag{3.4}$$

From the data, I know the probability of each respondent $p(y > y_{mid})$ (let's call it p). Therefore, I use the triangular functional form in the case where $y_{mid} < y < y_{max}$:

$$p = 1 - \frac{(y_{max} - y)^2}{(y_{max} - y_{min})(y_{max} - y_{mid})} \tag{3.5}$$

If I isolate the y, we have :

$$(y_{max} - y_{min})(y_{max} - y_{mid}) - p(y_{max} - y_{min})(y_{max} - y_{mid}) = (y_{max} - y)^2 \tag{3.6}$$

And finally we have :

$$y = y_{max} - \sqrt{(y_{max} - y_{min})(y_{max} - y_{mid}) - p(y_{max} - y_{min})(y_{max} - y_{mid})} \tag{3.7}$$

Equation 3.7 captures the expected earnings of each respondent in ten years if they are waged- or self-employed and they attended university.

Then, the expected gross returns to university is computed by taking the difference between expected earnings of university and expected earnings of having no post-secondary education : $g^{uni} = y^{uni} - y^{nopse}$. This measurement of returns to education is conditional to obtain a waged- or self-employed job. In order to capture the perceived risk of unemployment after having attended university, I use the probability of working as wage- or self-employed in ten years $q \in (0, 10)$. I think that individuals consider the returns to education by taking into account the probability of being employed. Indeed, even where the returns to university are very high, if the probability of employment is very low, the individuals will probably not choose to apply to and attend university. This is the reason why I compute the returns to education unconditional to employment by multiplying the returns to education with the probability of employment : $return_{uncondi} = return_{condi} * q$. Finally, I proxy the perceived risk of earnings by a measure of the spread of $y : w = y_{max} - y_{min}$. Information for these variables was collected from 2011 to 2013. To compare the values, the deflated variables are considered and expressed at the 2018 value.

3.5 DESCRIPTIVE STATISTICS

This section presents some descriptive statistics in order to better understand the pattern of the main variables used in the analysis. The first part focuses on the two dependent variables, then light is shed on the main interest variables : expected returns, academic ability and perceived ability. The last part of the section focuses on individual and family characteristics, which are used as control variables. In order to better understand the context of tertiary education in Ghana, statistics are presented for all the post-secondary institutions that a student may choose after SHS. However, the analysis (in Sections 6 and 7) focuses on the choice of application to and attendance at university only. Indeed, university is the post-secondary institution that has the most applications and the most attendance as well. Moreover, university is the most generalist tertiary institution.

3.5.1 Attendance and Application

First, I look at the proportion of respondents that applied to and attend a post-secondary institution and which institution they chose. Table 3.4 reports that 30% of the sample never apply to a post-secondary institution and half of the sample has never attended a post-secondary institution. The highest proportions of application are for university and nursing, with 30% and 21% of application respectively. The preference for university persists at the step of the attendance, with 18.4% of individuals having attended or currently attending a university, while teaching and nursing follow

with 10%. The data shows gender differences. If the first choice of post-secondary institution is university for men, women prefer nursing. Boys have a lower probability to apply to and attend a post-secondary institution than girls. About 55.75% of the sample have only ever attended one type of post-secondary institution, 4.02% have attended two different types and 0.05% have attended three.

Second, I present the probability to attend an institution conditional on the application to this institution. Results change slightly : the most attended institution is not university anymore but polytechnic. Indeed, 74.6% of the persons who applied to polytechnic attended it whereas it is the case for 60.7% of students for applied to university. Regarding teaching and nursing, it is likely that these are “second best" choices as around 50% of students who applied to these institutions finally attend it. Interestingly, men are more likely than women to attend university or polytechnic after applying to them. On the other side, women are more likely to attend teaching and nursing schools after having applied to them compared to men.

3.5.2 Expectations on Labor Market

Table 3.5 shows different variables of expectations on the labor market. Interestingly, the expected returns on attending university conditional to employment are higher than other types of post-secondary institution. However, the probability of employment is lower for university (83%) than for nursing (85%) and teacher training (84%), though this difference is very small. Respondents estimate that the probability to find a job is largely higher for people who attended a post-secondary institution than for people who did not attend any post-secondary institution (54% probability to find a job). The range of variation of earnings is the highest for people who attended university, following by nursing. On average, men have better expectations of the labor market in terms of both probability of employment and earnings than women, but the difference is thin.

In order to compare the actual earnings to the real expected earnings, I use the weekly earnings from the primary job of the respondents and convert and deflate it into a real monthly income. Table 3.6 shows that the expected earnings are largely overestimated by the respondents. This difference can be explained in part by the fact that the survey refers to the expected earnings in ten years, whereas the actual earnings were collected between 2011 and 2018. However, the comparison gives an indication of the information that the SHS students have on the labor market. It reveals that the students from public high schools in Ghana overestimate their future earnings. The evidence is similar to what Avery and Kane (2004) find in Boston, USA. Several reasons can explain this phenomenon. First, students have unrealistic expectations on the returns to education because they do not have good and accurate information about the labor market. They can also base their expectations on the situation of their relatives, which is not representative of the actual situation on the labor market.

TABLEAU 3.4 Attendance and Application to a Post-Secondary Institution

	University	Polytechnic	Teacher	Nursing	No PSE
Application (=1 if applied)					
Whole	30.02	12.17	19.37	20.91	30.45
	(45.84)	(33.31)	(39.53)	(40.67)	(46.02)
Men	34.42	14.24	19.67	12.99	31.21
	(47.52)	(34.96)	(39.76)	(33.63)	(46.34)
Women	23.39	10.37	18.92	32.85	29.29
	(42.35)	(30.50)	(39.18)	(46.98)	(45.53)
Attendance (=1 if attended)					
Whole	18.38	9.59	10.50	10.25	51.93
	(38.73)	(29.46)	(30.66)	(30.35)	(49.97)
Men	21.02	11.09	10.39	5.57	52.58
	(40.76)	(31.41)	(30.52)	(22.94)	(49.94)
Women	14.38	7.34	10.67	17.33	50.95
	(35.10)	(26.09)	(30.89)	(37.87)	(50.01)
Attendance conditional on application					
(=1 if attended, =0 if applied but not attended, missing if not applied)					
Whole	60.70	74.58	52.65	48.48	99.41
	(48.86)	(43.59)	(49.97)	(50.01)	(7.69)
Men	60.90	76.76	51.02	42.47	99.36
	(48.82)	(42.31)	(50.05)	(49.52)	(7.99)
Women	60.19	70.07	55.20	52.07	99.48
	(49.03)	(45.96)	(49.82)	(50.01)	(7.18)

Note : Indicates the percentage of SHS-graduated respondents who have ever applied and/or attended (or are currently attending) each type of post-secondary institution. Standard deviations are in parentheses. Computed with one observation per respondent ($N_{whole} = 3314$; $N_{men} = 1993$; $N_{women} = 1321$ for application and attendance unconditional; $N_{whole} = 995$; $N_{men} = 686$; $N_{women} = 309$ for attendance conditional on application). 'No PSE' means 'No Post-Secondary Education'. Anne Duplantier (2021)

TABLEAU 3.5 Expectations about the Labor Market

	No PSE	University	Polytechnic	Teacher	Nursing
		Whole sample (N=3314)			
Return educ.	-	3241.49	1976.06	1635.93	2277.02
(condi empl.)	-	(2815.31)	(1938.36)	(1582.66)	(1891.40)
Prob. work.	54.16	83.12	76.95	84.46	85.31
	(23.84)	(16.48)	(17.96)	(17.77)	(17.32)
Return educ.	-	2738.63	1548.65	1400.54	1981.04
(uncondi empl.)	-	(2522.13)	(1610.37)	(1440.65)	(1772.35)
Max-min earn.	591.05	2112.90	1404.92	1250.84	1502.81
	(538.28)	(2189.62)	(1374.78)	(1181.80)	(1442.27)
		Men (N=1993)			
Return educ.	-	3374.07	2030.72	1647.02	2202.46
(condi empl.)	-	(3012.03)	(2004.79)	(1642.32)	(1866.65)
Prob. work.	54.15	83.25	77.33	84.83	84.92
	(23.83)	(16.37)	(17.52)	(17.37)	(17.08)
Return educ.	-	2855.54	1592.93	1408.90	1908.46
(uncondi empl.)	-	(2707.01)	(1648.09)	(1476.54)	(1736.04)
Max-min earn.	6602.37	2207.03	1446.68	1263.09	1495.64
	(541.17)	(2345.38)	(1414.22)	(1187.80)	(1422.67)
		Women (N=1321)			
Return educ.	-	3041.46	1895.40	1619.27	2393.09
(condi empl.)	-	(2476.98)	(1833.74)	(1489.11)	(1924.30)
Prob. work.	54.18	82.92	76.38	83.92	85.92
	(23.85)	(16.66)	(18.60)	(18.34)	(17.68)
Return educ.	-	2562.24	1483.10	1387.96	2094.14
(uncondi empl.)	-	(2203.46)	(1551.18)	(1385.48)	(1822.44)
Max-min earn.	573.97	1970.89	1343.23	1232.47	1513.97
	(533.63)	(1923.16)	(1312.48)	(1172.96)	(1472.77)

Note : Standard deviations are in parentheses. The earnings are in real terms (value of 2018). 'No PSE' means 'No Post-Secondary Education'. Anne Duplantier (2021)

Finally, they could overestimate their own capacities - even if they have accurate information on labor market, students may think that they are better and will earn more than the average.

TABLEAU 3.6 Expected versus Actual Earnings

	All	No PSE	University	Polytechnic	Teacher	Nursing
Real exp. earnings	-	1195.57	4429.86	3193.98	2812.35	3487.96
	-	(945.03)	(3244.28)	(2358.41)	(1984.75)	(2382.33)
	-	[4203]	[3936]	[4101]	[4135]	[4145]
Real actual earnings	565.20	527.71	785.77	713.51	1068.24	978.66
	(590.36)	(615.38)	(652.67)	(638.85)	(378.59)	(725.28)
	[4796]	[2557]	[423]	[177]	[242]	[181]
Real actual earnings	843.99	-	784.677	697.72	1075.97	982.41
(graduated only)	(604.74)	-	(653.23)	(634.93)	(351.25)	(734.95)
	[1164]	-	[403]	[165]	[232]	[169]

Note : The database is used in its panel form, i.e. there are several observations per individual. Standard deviations are in parentheses and the number of observations is between brackets. 'No PSE' means 'No Post-Secondary Education'. Anne Duplantier (2021)

3.5.3 Expectations on Abilities

Table 3.7 presents the different measurements of ability. The academic ability is measured by the score that the student obtained during the final exam of the SHS i.e. the WASSCE. In Ghana, the WASSCE score varies from 6, the best score, to 54, the worst score. The average score for the whole sample is 24, corresponding to the minimal score required to be accepted at university. [26] On average, respondents who attended university have better scores than those who attended other types of post-secondary institution. Not surprisingly, the respondents who did not attend any post-secondary institution have the worst score.

Expectations on WASSCE score allow a comparison between expected and actual WASSCE score. As for expectations on the labor market, students tend to overestimate their academic abilities. They expect on average a score of 10.5 while the average score is 24. Similarly, respondents who attended university have on average higher ex-ante expectations on their WASSCE results than others. Respondents who did not attend any post-secondary institutions have the lowest expected score.

26. According to the website of the 'University of Ghana', minimum/general entry requirements are as follows : "An applicant for admission to a degree program must have at least credits (A1-C6 in WASSCE) in English, Core Mathematics and Integrated Science (for Science-related programs) or Social Studies (for non-Science-related programs) and three elective subjects in Science for applicants applying to Science- or Agriculture-related disciplines or three elective subjects in General Arts/Business for applicants applying to non-Science-related disciplines, with the total aggregate not exceeding 24. In addition, Science applicants should have at least a grade C6 in WASSCE in Social Studies and non-Science applicants should also have at least a grade C6 in WASSCE in Integrated Science." http://admission.ug.edu.gh/applying/content/20172018-entry-requirements-ghana-wassce

The ex-ante perceived ability to be admitted into university is very high. The respondents estimate on average that the likelihood to be accepted into university with their first WASSCE score [27] is 79%. Among the respondents who attended university, the likelihood is the highest (82%), whereas the respondents who did not attend a post-secondary institution have the lowest (77%).

Finally, the data brings information on the expected likelihood to afford attending university. On average the perceived academic ability is higher than the perceived financial ability. Respondents estimate on average that they have a probability of 68% to be able to finance their studies in university. There are differences according to the gender. Boys have a higher perceived probability to be accepted into university than girls. However, girls report a higher expected probability to be able to afford attendance at university than boys. In general, boys and girls expect similar WASSCE results. However, on average boys have better results.

TABLEAU 3.7 Perceived and Actual Ability

	All	Men	Women	Uni.	Poly.	Teach.	Nurs.	No PSE
Actual av. WASSCE score	26.18	25.47	27.26	20.22	28.15	24.52	25.41	29.06
	(9.09)	(9.35)	(8.59)	(8.46)	(7.07)	(6.76)	(7.39)	(8.82)
Exp. av. WASSCE score	10.14	10.11	10.18	8.65	10.16	10.28	9.67	10.88
	(3.32)	(3.47)	(3.07)	(2.54)	(2.64)	(2.80)	(2.77)	(3.63)
Prob. be admitted uni. (%)	78.87	79.58	77.79	82.08	79.31	79.65	77.72	77.47
	(18.27)	(18.01)	(18.64)	(15.58)	(18.22)	(18.73)	(18.81)	(19.07)
Prob. afford uni. (%)	68.04	67.74	68.52	71.63	68.96	63.02	72.11	66.34
	(20.22)	(20.28)	(20.13)	(20.00)	(20.29)	(19.70)	(20.09)	(20.13)
Observations	1323	803	520	312	87	116	123	687

Note : One observation per respondent. Standard deviations are in parentheses. WASSCE scores vary between 6 (the best) and 54 (the worst). Anne Duplantier (2021)

3.5.4 Individuals Characteristics

Table 3.8 describes the individual and familial characteristics of the sample used in the analysis. About 60% of the sample are men of 19 years old on average. More than half of the sample are born in a rural area and from a household with four children on average. Amongst the fathers of respondents, 41.5% have at least a SHS level of education whereas 19.5% have no education. Regarding the mothers, 29.7% have no education and 20.3% have at least a SHS level of education. For both parents, the highest level of education is JHS, meaning that on average this generation of students are more educated than their parents. Moreover, as seen in previous chapters, mother's respondent is in general less educated than father.

27. In Ghana, students first take the WASSCE at the end of the SHS and then can take it again to improve their score.

TABLEAU 3.8 Individual and Familial Characteristics

	Mean
Age (years)	19.61
	(2.59)
Male (%)	60.14
	(48.97)
Born in rural area (%)	54.01
	(49.84)
Num of siblings	4.10
	(2.26)
Father educ (%)	
None	19.40
Primary	4.35
JHS	34.67
SHS	22.27
Post-secondary	19.31
Mother educ (%)	
None	29.72
Primary	9.96
JHS	40.01
SHS	14.06
Post-secondary	6.25
Region of SHS (%)	
Ashanti	16.17
Brong Ahafo	11.35
Central	10.74
Eastern	14.67
Greater Accra	8.00
Northern	6.91
Upper East	3.56
Upper West	3.29
Volta	16.87
Western	8.45
Cohort (%)	
2005	10.05
2006	12.85
2007	18.11
2009	22.78
2010	21.42
2011	14.79
Observations	3314

Note : Standard deviations are in parentheses. Anne Duplantier (2021)

3.6 EMPIRICAL METHODOLOGY

This section explains the empirical methodology used in this chapter. The first part presents the estimation, the variables and the expected effects. The second part sheds light on the threats to the identification strategy. The last part discusses the other estimation methods, as well as their advantages and limits.

3.6.1 Estimation Strategy

As a first step, the links between expectations on the labor market and the returns to education are examined with a linear probability model. As it is done in the literature, expectations on the labor market are captured by two variables : the expected returns to education unconditional to employment and the perceived risk of earnings. As explained previously, the analysis focuses on the choice of pursuing into university after graduating from SHS. Two types of education outcomes are used - application and attendance conditional on application - so there are two dependent variables. Estimations are as follows :

$$\begin{aligned} appl\_uni = {} & \alpha_1 exp\_return\_uncondi + \alpha_2 exp\_maxmin \\ & + \delta control\_var + \lambda_1 cohort\_FE + \lambda_2 region\_FE + \epsilon_{clusterSHS} \end{aligned} \tag{3.8}$$

$$\begin{aligned} att\_uni = {} & \beta_1 exp\_return\_uncondi + \beta_2 exp\_maxmin \\ & + \varphi control\_var + \gamma_1 cohort\_FE + \gamma_2 region\_FE + \mu_{clusterSHS} \end{aligned} \tag{3.9}$$

With 3.9 identified only for observations for which $appl\_uni = 1$.

Then, the paper contributes to the literature by considering perceived and academic ability as other determinants of education choices.

$$\begin{aligned} appl\_uni = {} & \alpha_1 exp\_return\_uncondi + \alpha_2 exp\_maxmin + \alpha_3 wasscescore \\ & + \alpha_4 perceived\_prob\_admitted\_uni + \alpha_5 perceived\_prob\_afford\_uni \\ & + \delta control\_var + \lambda_1 cohort\_FE + \lambda_2 region\_FE + \epsilon_{clusterSHS} \end{aligned} \tag{3.10}$$

$$\begin{aligned} att\_uni = {} & \beta_1 exp\_return\_uncondi + \beta_2 exp\_maxmin + \beta_3 wasscescore \\ & + \beta_4 perceived\_prob\_admitted\_uni + \beta_5 perceived\_prob\_afford\_uni \\ & + \varphi control\_var + \gamma_1 cohort\_FE + \gamma_2 region\_FE + \mu_{clusterSHS} \end{aligned} \tag{3.11}$$

With 3.11 identified only for observations for which $appl\_uni = 1$.

Definition of the Variables

The dependent variable $appl\_uni$ is a dummy equal to 1 if the respondent applied to university and 0 if not, and att_uni is a dummy equal to 1 if the respondent attended university and 0 if not. Notice that the variable att_uni is conditional to having applied to university. Thus, att_uni is identified only for observations for which $appl\_uni = 1$. The explicative variable $exp\_return\_uncondi$ comes from a several-steps procedure. Details on how the expected returns to every level of education are computed were previously explained in Section 4. The next step is to take the difference between expected earnings of university and expected earnings of having no post-secondary education. Then the expected returns to university are multiplied by the probability of working to obtain the expected returns to university unconditional to employment. The perceived risk of earnings is captured by the variable exp_maxmin which is measured with the spread between the maximum expected earnings and the minimum expected earnings. Information for these two variables was collected from 2011 to 2013, so the variables are deflated and expressed at the 2018 value.

The academic ability is measured by the variable $wasscescore$ which is the WASSCE score reversed and normalized from 0 to 100. The variable $perceived\_prob\_admitted\_uni$ captures the perceived academic ability to enter into university. The respondents were asked to reveal the probability that the WASSCE results will be good enough to enter into university. This variable varies from 0 to 10. Finally, the variable $perceived\_prob\_afford\_uni$ measures the perceived ability to afford university. The respondents were asked the probability they would be able to afford attending university as a regular student (i.e. non-fee paying student) if they were admitted. This variable varies from 0 to 10.

Finally, the estimation includes individual and family characteristics that are likely to influence education decisions, as discussed previously in Sections 3. The control variables include a variable equal to the age of the respondent and a dummy indicating the sex of the respondent. The variables capturing the education level of the parents take the form of dummy equal to 1 for each level of education and each parent : no education (which is the exclusion category), primary education level, JHS level, SHS level and post-secondary education level. The other control variables are the number of siblings and a dummy indicating whether the respondent is born in a rural or urban area. In order to control for regional and cohort effects, the estimation includes both regional and cohort fixed effects. Moreover, standard errors are clustered to account for the fact that students may have common and unobservable characteristics within the SHSs.

Expected results

Following the literature, [28] I expect a positive correlation between expected returns to education un-

28. For example, Attanasio and Kaufmann (2014), Jensen (2010) and Arcidiacono et al. (2012).

conditional to employment and the decision to pursue education after SHS. Regarding the variable that captures the range of earnings, the direction of the relationship will depend on the attitude toward risk of the students. If students are risk-averse, they will choose a type of post-secondary education that minimizes the range of earnings variation. But if they are not risk averse and are attracted by high earnings, the relationship can be positive.

I believe that both perceived and academic ability are positively related to the both decisions to apply and attend a post-secondary institution. First, in order to be admitted into university in Ghana, there are requirements and cutoffs based on the WASSCE score of the respondent. A higher WASSCE score is expected to be positively related to the probability to apply to university as well as the probability to attend university.

In addition to the academic ability, I believe that the perceived ability of the students can also have an impact on the decision to pursue schooling after SHS. If the student does not think that they are able to be admitted and afford university, they will neither apply to nor attend university. This explains why I believe that a higher perceived ability to be admitted into university and a higher perceived ability to afford university encourage the respondent to apply. Therefore, I am expecting β_1, β_2 and β_3 to be positive for application and attendance to university.

3.6.2 Threats to the Identification Strategy

A linear probability model (LPM) is employed to estimate the regressions previously presented. In particular, the estimator Ordinary Least Square (OLS) with clustered standard errors at the SHS level is used. Two main limitations are put forward in the literature against the use of OLS with binary dependent variable. First, OLS imposes heteroscedasticity when using a binary variable. However, this weakness is addressed by using clustered standard errors at the SHS level. Indeed, cluster robust standard errors are also heteroskedastic consistent (Cameron and Miller, 2015). Second, the weakness of OLS with binary variable is that the predicted probability is not constrained to the 0-1 interval. However, Horrace and Oaxaca (2006) find that if no or very few predicted probabilities lie outside the 0-1 interval, the LPM is expected to be unbiased and consistent.

Moreover, an increasing body of literature sees LPM as a suitable alternative to non-linear models such as probit or logit. Angrist (2001) reports that challenges of estimating binary dependent variables are not really different from estimating continuous variables. In that respect, Ashraf (2009) uses OLS, arguing that it functions similarly to probit and logit and has the advantage of imposing less structure on the data. For Friedman (2012), non-linear models require strong assumptions that lead to bias if assumptions are wrong. Moreover, Hippel (2015), Ashraf (2009) and Gomila (2019) highlight that the major asset of the LPM is its interpretability. Indeed, LPM allows an easier interpretation of coefficients than logistic and probit models. Gomila (2019) shows that LPM is “the

most powerful, flexible and the simplest strategy" in the presence of binary variables.

The other general limit to the OLS estimator is the endogeneity issue. Endogeneity may arise from reverse causality between the dependent variables and the variables of interest. Could expectations on self-ability and the labor market be influenced by the decision to apply to and attend a post-secondary institution ? Jensen (2010) highlights this endogeneity issue, giving the example of “sour grapes" : a respondent who really wishes to go to school but is constrained financially or by their abilities can underestimate the returns to education. Thanks to the time gap between the collection of expectations and education choice, the risk of the survey facing reverse causality issues is reduced. The information on expectations was collected when students were at school whereas the information on schooling decisions was collected once students left SHS. The schooling decisions of the students should therefore not have an impact on the expectations they had when attending SHS.

However, one cannot exclude a situation where a student already knows precisely their post-secondary education choice early in SHS and adapts their expectations following these choices. I believe that it is unlikely that a lot of SHS students know so early what they will do after SHS. Nevertheless, a student can already know that they are constrained because of their academic results or their financial situation. In such a context, it is possible to have a bias from constrained respondents who might underestimate their ex ante expectations. Thanks to the survey, I am able to check the profile of the respondents who have low and high expectations.

Table 3.9 shows statistically significant differences in the characteristics of individuals with high and low expectations. Students from the 90th percentile of expected returns have higher actual WASSCE scores than those from the 10th percentile. They estimate that they have a higher probability to be admitted into university and to be able to afford associate fees. The respondents from the 90th percentile are relatively younger than the ones from the 10th percentile. They come from smaller families. They have a lower probability to have parents with low education and more probability to have parents with at least an SHS level of education. If I look at the differences around the median, the differences in characteristics are less significant and smaller than for the 90th and 10th percentiles. In sum, I cannot exclude that constrained respondents may underestimate their expectations, and I should keep in mind this potential limitation.

The second reason why the OLS estimator can be biased is the measurement errors. As discussed in the literature review, despite some doubts on the use of expectations in economics, it has been shown that respondents understand well the concept of probability and expectations. To check this in this database, I presented statistics of probability practice in Section 4 and it seems very consistent. I do not see why the respondents would hide their real expectations. Furthermore, as expectations are what people think and estimate, I do not believe that expectations can be measu-

TABLEAU 3.9 Difference in Characteristics Between Individuals with High and Low Expectations

	(1)	(2)	(3)	(4)	(5)	(6)
Expected returns uncondi.	90th perc.	10th perc.	(1) - (2)	Sup. 50th	Inf. 50th	(4) - (5)
Exp. max-min	5707.47	825.93	3272.69***	3114.80	1133.83	1980.97***
	(197.31)	(39.46)	(201.67)	(58.83)	(16.81)	(61.19)
Actual wassce score	62.88	57.11	5.76***	60.34	58.81	1.53**
	(1.09)	(1.08)	(1.53)	(0.45)	(0.45)	(0.64)
Percvd prob. admitted uni	82.23	71.33	10.90***	80.57	75.16	5.41***
	(1.07)	(1.44)	(1.76)	(0.49)	(0.59)	(0.76)
Percvd prob. afford uni	68.99	62.98	6.01***	69.19	67.08	2.11**
	(1.27)	(1.36)	(1.87)	(0.59)	(0.59)	(0.83)
Age	18.96	19.87	-0.91***	19.49	19.86	-0.37***
	(0.13)	(0.13)	(0.18)	(0.06)	(0.06)	(0.08)
Male	64.70	58.13	6.57*	59.94	57.69	2.25
	-	-	(3.41)	-	-	(0.01)
Father - No education	13.60	21.07	-7.47***	17.78	20.87	3.10**
	(1.82)	(2.16)	(2.83)	(0.90)	(0.95)	(1.31)
Father - Primary education	3.68	4.77	-1.09	3.80	4.48	0.68
	(1.00)	(1.13)	(1.51)	(0.45)	(0.48)	(0.66)
Father - JHS education	28.89	34.83	-5.94*	33.41	36.99	-3.59**
	(2.41)	(2.53)	(3.49)	(1.11)	(1.13)	(1.58)
Father - SHS education	31.44	22.19	9.25***	24.99	18.96	6.02***
	(2.47)	(2.20)	(3.31)	(1.01)	(0.92)	(1.37)
Father - Post-sec education	22.37	17.13	5.24*	20.03	18.69	1.34
	(2.22)	(1.99)	(2.98)	(0.93)	(0.91)	(1.31)
Mother - No education	20.05	33.24	-13.19***	27.39	32.29	-4.90***
	(2.10)	(2.45)	(3.23)	(1.03)	(1.07)	(1.49)
Mother - Primary education	8.52	10.81	-2.29	10.11	9.39	0.72
	(1.46)	(1.62)	(2.18)	(0.70)	(0.67)	(0.97)
Mother - JHS education	43.41	37.30	6.11*	40.34	39.21	1.14
	(2.60)	(2.52)	(3.62)	(1.13)	(1.12)	(1.59)
Mother - SHS education	18.41	12.70	5.70**	14.93	13.03	1.89*
	(2.03)	(1.73)	(2.67)	(0.82)	(0.77)	(1.13)
Mother - Post-sec education	9.61	5.94	3.67*	7.22	6.06	1.15
	(1.55)	(1.23)	(1.97)	(0.60)	(0.55)	(0.81)
Number of siblings	3.83	4.11	-0.28*	3.94	4.11	-0.16**
	(10.14)	(10.12)	(14.33)	(0.05)	(0.05)	(0.07)
Born in rural area	43.51	55.64	-12.13***	49.16	56.62	-0.07***
	-	-	(3.57)	-	-	(1.59)
Observations	381	382	-	1914	1914	-

Note : *** p<0.01, ** p<0.05, * p<0.1. Standard deviations are in parentheses. Anne Duplantier (2021)

red with errors. Beyond the question of understanding, the other potential issues of measurement are the calculation of expected earnings. I make an assumption on the functional form (triangular distribution) that may produce some approximation and measurement errors.

Finally, endogeneity may also arise from omitted variables bias. There are several variables that could have been included in the estimation if they were available. First, the family's wealth is known to be positively related to the education choices of the student, as highlighted by the World Bank's World Development Report 2018. [29] But I have good reasons to think that the income level of the household can also influence positively the respondent's expectations of their ability and the labor market. Furthermore, motivation and intellectual ability are also positively related to the education choices, and, at the same time, to the expectations of the respondent. These three potential omitted variables can bias the estimations in the same positive way. There is, then, a risk of overestimating the effect with the OLS estimator. One way to avoid this bias is to use proxies to these variables, as I cannot measure it directly. Indeed, I can proxy the family's income with the both variables of parent's education level and the perceived probability to be able to finance university. The more the parents have a high level of education and the more the perceived ability of the respondent to be able to afford university is high, the more the income of the family should be high. Regarding the motivation and the intellectual capacity, the WASSCE score of the respondent could capture some part of the effects.

3.6.3 Discussion on Alternative Estimation Methods

In order to face the endogeneity issue from the OLS estimation, other estimation methods are available. One solution is to find an exogenous determinant of expectations that could be used as an instrumental variable. The latter should be related to the expectation variable but not to the education decisions except through the expectations effect. But it is difficult to find the ideal candidate as economic literature has not shed light on the determinants of expectations.

Another way to address this issue is the regression discontinuity design (RDD) that uses cutoff as random assignation. As it can be considered that individuals just below a cutoff have globally the same characteristics that those just above, one could therefore compare these two groups of individuals as differing only according to the cutoff. However, several conditions should be met to ensure the validity of the method. Lee and Lemieux (2010) highlight that the RDD is invalid if the assignment is not exogenous. This happens if the assignment variable that determines whether an individual is below or above the cutoff can be manipulated by respondents. Sequeira et al. (2016) use a fuzzy RDD to identify a causal relation between the fellowship award and perceived returns to education. In the context of this chapter, to conduct a RDD, I could use as a cutoff the WASSCE

29. The report is called 'Learning to Realize Education's Promise' (WorldBank, 2018).

score allowing respondents to be admitted into university. However, in Ghana this cutoff is well known by students as it is announced on the website of the post-secondary institutions. Moreover, students are able to take the WASSCE as many times as they need in order to obtain the desired score. Here the assignment is completely manipulable by participants and is not exogenous. The approach is therefore invalid and not applicable in this case.

Another method frequently used is the propensity score matching (PSM). This method allows researchers to relax the assumption of functional form and to build a comparison group. The counterfactual is built by associating an individual that received the treatment to another with similar observables characteristics but who did not receive the treatment. For this to be valid, we need to assume that there is no selection under non-observable characteristics. The potential problems of omitted variables from the OLS estimation will not be solved with this method. Moreover, the PSM methodology requires a treatment to evaluate, as well as a treatment group that received the treatment and a control group that did not. In this chapter, I cannot identify the expectations of the respondents as a treatment. Indeed, every respondent had formulated expectations (either high or low). Furthermore, I cannot separate the sample in two groups and distinguish a treatment.

While the original database is longitudinal, an estimation in panel does not work either. The reason is that the interest outcomes - application and attendance to university - do not vary over time. These variables are built from the later waves (once respondents had graduated from SHS), based on the choices they made. If the respondent attended university once (in one wave, but not in the following wave), the variable will always consider it as though they did.

Finally, it worth mentioning the identification through heteroscedasticity method. This method provides an estimator for linear regression models including endogenous regressor but no instrument variables are available (Lewbel, 2012). It uses the natural heteroscedasticity of the data and the random shocks to obtain a role similar to standard instruments. For the identification, the method uses the explicative variables that are uncorrelated with the covariance of heteroscedastic errors. According to Lewbel (2012), such variables are available in many models in which endogeneity comes for example from measurement errors. The advantage of this method over instrumental variables is that it does not rely on exclusion restrictions. The main limit is to give noisier and less reliable estimates than identification based on standard exclusion restrictions. Hogan and Rigobon (2002) use the heteroscedasticity method to estimate the returns to education. Their results are very close to the OLS estimators and support the conclusion of Griliches (1977) who suggests that OLS is a pertinent method for estimating the returns to schooling despite its limits.

Having reviewed the potential methods, it seems that despite the limits of the OLS estimator, it is in this case the best approach. Notice that the probit and logit estimators will be used in a robustness analysis.

3.7 RESULTS

This section presents the results of the analysis. First, the primary results are discussed, then the analysis is deepened with heterogeneity effects. Finally, a sensitivity analysis is conducted in order to test the robustness of the results.

3.7.1 Primary results

The first findings presented in Table 3.10 show that the expected returns to university unconditional to employment are significantly and positively related to the decision to apply to university after graduating from SHS (Column (1)). When expected returns to university increase by 100 GHS, the likelihood to apply to university increases by 1.9 units. In Column (2), the difference between maximum and minimum expected returns to university (proxying the perceived risk of earnings) is included in the estimation. The coefficient of the expected returns remains relatively stable and significant, but the perceived risk of earnings does not have a significant effect on the probability to apply to university. After adding control variables in Column (3), results stay the same. The expected returns to university are still significantly related to the likelihood of application to university, but the magnitude decreases slightly (from 0.019 to 0.013 units).

Results in Column (3) suggest that individual characteristics play a role in the decision to apply to university. First, younger students have a higher probability to apply to university. If age increases by 1 year, the likelihood of applying to university decreases by 0.03 units. As fixed effects at the cohort level are included, this is not a cohort effect. It suggests that within the same cohort, the younger students have a higher probability to apply to university. This result can be explained by the fact that younger students are probably the most motivated compared to students that repeated a class or entered school at later than the normal age. In addition, older students may have fewer incentives to pursue education as their opportunity cost is higher. Second, boys have more chances to apply to university than girls. Compared to a female student, a male has a higher probability of 0.14 to apply to university. This can be explained by factors that limit more girls in their education decisions, such as financial constraints or lack of confidence and ambition. Differentiated effects by gender will be addressed later in the analysis.

Students with more siblings have a significantly lower probability to apply to university (but the coefficient is almost null). Finally, the findings show that parents' education level is positively linked to the decision to apply to university. Indeed, a student whose father has a post-secondary education level has a higher probability (of 0.073) to apply to university than a student whose father has no education. Similarly, a student whose mother has a primary education level has a higher probability (of 0.06) to apply to university than a student whose mother has no education. A

TABLEAU 3.10 University Decision and Expectations about the Labor Market

	(1)	(2)	(3)	(4)	(5)	(6)
		Application			Attendance	
Exp. ret. uncondi. empl.	0.019***	0.016***	0.013**	0.005	0.004	0.002
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Exp. max-min		0.015	0.011		0.035	0.032
		(0.013)	(0.012)		(0.024)	(0.024)
Age			-0.032***			-0.024**
			(0.005)			(0.010)
Male			0.135***			0.023
			(0.022)			(0.032)
Father primary educ			-0.039			-0.042
			(0.044)			(0.096)
Father JHS educ			-0.029			-0.015
			(0.029)			(0.051)
Father SHS educ			-0.018			-0.006
			(0.031)			(0.061)
Father post-sec educ			0.073**			-0.014
			(0.031)			(0.063)
Mother primary educ			0.060**			-0.010
			(0.030)			(0.070)
Mother JHS educ			0.022			0.026
			(0.023)			(0.048)
Mother SHS educ			0.033			0.111**
			(0.031)			(0.054)
Mother post-sed educ			0.103**			-0.025
			(0.042)			(0.067)
Number of siblings			-0.008**			-0.000
			(0.003)			(0.008)
Born in rural area			-0.019			-0.001
			(0.017)			(0.037)
Constant	0.230***	0.128	0.816***	0.400***	0.168	0.705**
	(0.043)	(0.099)	(0.133)	(0.073)	(0.185)	(0.300)
Region fixed effects						
Cohort fixed effects						
Observations	3314	3314	3314	995	995	995
R square	0.047	0.047	0.096	0.047	0.049	0.062
Adjusted R square	0.042	0.042	0.088	0.032	0.033	0.035
Average interest var.	0.300	0.300	0.300	0.607	0.607	0.607

Note : *** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at SHS level and are in parentheses. Anne Duplantier (2021)

significant and stronger effect is found if the mother has a post-secondary education level. Having a mother with a higher education level increase the probability by 0.103, compared to a student whose mother has no education. Being born in a rural area does not have a significant impact on the decision to apply to university than being born in an urban zone. This suggests that there is no spatial discrimination in the application to university in this sample of students.

For the attendance at university (Columns (4)-(6)), the relationship between expected returns to university and decision is not statistically significant. In Column (5), the perceived risk of earnings is added in the estimation and, as for expected returns, it has no statistically significant relationship with the probability to attend university. Column (6) reports that some individual characteristics are significantly related to the decision to attend university (but there are fewer than for the application to university). As for the application to university, younger students are more likely to attend university, though the magnitude is slightly smaller than for application. Moreover, respondents whose mother has a secondary education level have a higher probability to attend university than respondents with a non-educated mother. Having a mother with a SHS education level increase the probability by 0.111, compared to a student whose mother has no education. Interestingly, this is the determinant that plays the bigger role. This suggests that once the student has applied to university, having an educated mother is one of the key determinant for attending university.

As a second step, the variables of actual and perceived ability are included in the estimations and reported in Table 3.11. For both attendance and application to university, the expected returns are not statistically significant. Moreover, the perceived risk of earnings is not significantly related to the decision to apply and attend university. Interestingly, the ability variables seem to be more significantly linked to the decision to apply to university than the expectations of the labor market after university. The actual academic ability (proxied by the WASSCE score) is significantly and positively related to both application and attendance. However, magnitudes of coefficients are close to zero. If WASSCE score increases of 1, the likelihood of application to university increases by 0.009 units and the likelihood of attendance increases by 0.006 units (Columns (4) and (8)). Unlike the actual academic ability, the perceived academic ability has no significant relationship with the decision to pursue schooling in university. This contrasts to Wiswall and Zafar (2014), who find that beliefs about own ability is a determinant of the decision of the college major. However, the perceived probability to afford university is significantly and positively related to application (although coefficients are close to zero). The perceived probability to afford university has a lower magnitude than the WASSCE score.

Adding the variables of ability does not change strongly the relationship of the individual characteristics with education decision (Columns (4) and (8)). The younger students have a higher probability to apply to university and attend university. The magnitude of the age coefficient is

TABLEAU 3.11 University Decision and Expectations about the Labor Market and Ability

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Application				Attendance			
Exp. ret. uncondi. empl.	0.004	-0.003	-0.006	-0.007	0.001	-0.004	-0.006	-0.007
	(0.009)	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)
Exp. max-min	0.017	0.019	0.020	0.018	0.009	0.009	0.009	-0.001
	(0.017)	(0.017)	(0.017)	(0.017)	(0.033)	(0.033)	(0.033)	(0.034)
Actual wassce score	0.010***	0.010***	0.010***	0.009***	0.007***	0.007***	0.007***	0.006***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Percvd prob. enter uni		0.001*	0.001	0.001		0.001	0.001	0.001
		(0.001)	(0.001)	(0.001)		(0.002)	(0.002)	(0.002)
Percvd prob. finance uni			0.002***	0.001***			-0.000	-0.000
			(0.000)	(0.000)			(0.001)	(0.001)
Age				-0.014*				-0.038**
				(0.008)				(0.018)
Male				0.108***				0.006
				(0.030)				(0.049)
Father primary educ				-0.067				-0.034
				(0.069)				(0.140)
Father JHS educ				-0.113**				-0.079
				(0.049)				(0.085)
Father SHS educ				-0.075				0.025
				(0.048)				(0.085)
Father post-sec educ				-0.034				-0.016
				(0.048)				(0.091)
Mother primary educ				0.027				-0.017
				(0.048)				(0.112)
Mother JHS educ				0.024				-0.018
				(0.042)				(0.075)
Mother SHS educ				-0.009				0.060
				(0.047)				(0.083)
Mother post-sed educ				0.035				-0.165*
				(0.061)				(0.096)
Number of siblings				-0.007				0.006
				(0.004)				(0.011)
Born in rural area				-0.023				-0.034
				(0.029)				(0.048)
Constant	-0.354***	-0.465***	-0.527***	-0.161	0.116	0.053	0.053	0.896*
	(0.131)	(0.149)	(0.147)	(0.235)	(0.247)	(0.280)	(0.287)	(0.461)
Region fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1323	1323	1323	1323	492	492	492	492
R square	0.176	0.179	0.182	0.201	0.076	0.076	0.076	0.112
Adjusted R square	0.167	0.169	0.172	0.184	0.049	0.047	0.045	0.058
Average interest var.	0.372	0.372	0.372	0.372	0.632	0.632	0.632	0.632

Note : *** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at SHS level and are in parentheses. Anne Duplantier (2021)

slightly smaller for the application compared to Table 3.10 (0.014 versus 0.032 units) and the difference is statistically significant (at 1% level). Regarding the attendance, the age coefficient is higher than compared to Table 3.10 (0.038 versus 0.024 units), but the difference is not statistically significant. Being a male is still related to a significant higher (of 0.112) chance to make an application to university. The gender effect is lower (0.068 units) but significant and positive for the decision to attend university. Students with more siblings are less likely to apply to university. Having one additional sibling decreases the likelihood of application to university by 0.009 units but there are no significant effects on the probability to attend university. Surprisingly, the students whose father has a JHS education level have less chance to apply to university than students whose father has no education at all. The other levels of education of both father and mother are not significantly related to the decisions to apply to and attend university.

3.7.2 Heterogeneity effects

In this section, differences by gender (Tables 3.12 and 3.13) and academic levels (Tables 3.14 and 3.15) are explored.

Table 3.12 presents the results for the decision to apply by splitting the sample between men and women. [30] Columns (1) and (2) report results with the expected returns to university only. Then, the perceived risk of earnings is included (Columns (3) and (4)), followed by the variables of academic ability, perceived academic ability and the perceived financial capacity in Columns (5) and (6). Finally, Columns (7) and (8) show results for the whole model (including individual characteristics).

First, we observe that the relationship between the expected returns to university and the probability to apply to university does not vary by gender : the link is not statistically significant for neither men nor women, except for women in Column (2) when there is no other variable (significant at level 10%). Second, the perceived risk of earnings is not statistically significant for boys (Column (5)), but is significant at 5% and positive for girls (Column (6)). However, this difference is not statistically significant. When the gap between maximum and minimum expected returns to university increases by 100 GHS, the likelihood to apply to university increases by 4.3 units for women. However, the significance level decreases to 10% when adding control variables (Column (8)) while the magnitude does not change.

Third, the actual WASSCE score of both women and men is significantly and positively related to the decision to apply to university (even after adding control variables). The magnitude is higher for male students : 0.011 units for boys versus 0.007 units for girls. This difference between the

30. The same analysis was done by using interaction variables by gender and the results are similar. The method of splitting sample is preferred because it allows all the variables to vary according to the gender instead of the interest variable only.

TABLEAU 3.12 University Application by Gender

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Men	Women	Men	Women	Men	Women	Men	Women
Exp. ret. uncondi. empl.	0.012	0.024*	0.006	0.016	-0.017	0.002	-0.016	0.001
	(0.011)	(0.014)	(0.013)	(0.015)	(0.012)	(0.014)	(0.012)	(0.014)
Exp. max-min			0.030	0.046*	0.018	0.043**	0.013	0.042*
			(0.025)	(0.025)	(0.021)	(0.022)	(0.021)	(0.021)
Actual wassce score					0.011***	0.007***	0.011***	0.007***
					(0.001)	(0.001)	(0.001)	(0.001)
Percvd prob. admitted uni					0.000	0.001	0.000	0.001
					(0.001)	(0.001)	(0.001)	(0.001)
Percvd prob. afford uni					0.002**	0.002***	0.001*	0.002***
					(0.001)	(0.001)	(0.001)	(0.001)
Age							-0.015*	-0.014
							(0.009)	(0.013)
Father primary educ							0.004	-0.210*
							(0.092)	(0.125)
Father JHS educ							-0.125*	-0.102
							(0.064)	(0.087)
Father SHS educ							-0.058	-0.108
							(0.060)	(0.085)
Father post-sec educ							0.043	-0.127
							(0.068)	(0.083)
Mother primary educ							0.012	0.060
							(0.068)	(0.073)
Mother JHS educ							0.036	0.032
							(0.052)	(0.066)
Mother SHS educ							-0.005	0.013
							(0.068)	(0.062)
Mother post-sec educ							0.033	0.043
							(0.090)	(0.080)
Number of siblings							-0.009	-0.005
							(0.007)	(0.006)
Born in rural area							-0.000	-0.024
							(0.032)	(0.038)
Constant	0.359***	0.121	0.185	-0.166	-0.477**	-0.713***	-0.072	-0.364
	(0.116)	(0.133)	(0.188)	(0.189)	(0.182)	(0.193)	(0.286)	(0.336)
Region fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	803	520	803	520	803	520	803	520
R-squared	0.023	0.084	0.025	0.090	0.203	0.185	0.208	0.191
Adjusted R-squared	0.007	0.063	0.008	0.067	0.186	0.159	0.185	0.156
Average uni appl.	0.426	0.288	0.426	0.288	0.426	0.288	0.426	0.288

Note : *** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at SHS level and are in parentheses. Anne Duplantier (2021)

two coefficients is statistically significant (the p-value of the test equals to 0.0095). This suggests that the boys' decision to apply to university is more strongly related to their WASSCE score than it is for the girls. Fourth, the perceived probability to be admitted to university is not significantly different according to the gender. For both men and women, this factor is not significantly related to the decision to apply to university.

Finally, the perceived probability to afford university is more significantly related to women's probability of application. The relationship is significant at 10% level for men, versus 1% for women. Moreover, the magnitude is higher for women (0.002) than men (0.001). However, the difference in coefficients between boys and girls is not statistically significant. The other individual characteristics are not significantly (at a 5% level of error risk or less) linked with the decision to apply to university, and there is no significant difference between men and women.

Table 3.13 presents the results of the same analysis differentiated by gender but for the decision to attend university. The only statistically significant difference between men and women in the relationship of university attendance and expectations is the WASSCE score. The WASSCE score is significantly (at 1% level) related to the likelihood of attendance for both women and men. However, the scale is almost two times stronger for male students than female students, even though the magnitude is almost null.

Interestingly, the perceived capacity to afford university fees is no more significant for the women's decision to attend university compared to their decision to apply to university. Financial ability seems to be important for the women's application step, but not for attendance. This suggests that women probably limit themselves before applying : the more they feel they cannot afford university fees, the less they apply to university. But the perceived ability to finance university has no effect on the probability to attend.

Tables 3.14 and 3.15 show the results of the analysis differentiated by academic levels. Indeed, students with better academic results might have different behavior facing the schooling decision compared to students with lower score. Two groups of students are considered according to their WASSCE score : students with better academic ability, i.e. having a WASSCE score superior to the sample mean score, [31] and students with lower academic ability, i.e. having a WASSCE score inferior to the mean.

As for the gender effect results, variables are included step by step in the regression. Table 3.14 presents the results for the decision to apply to university. For both groups of students, the expected returns to university are not significantly related to the likelihood of application. The perceived risk of earnings appears to be positively and significantly (at 5% level) linked to better students'

31. The average WASSCE score equals 62.5 while the median WASSCE score equals 64.5. As both are quite close, we take the average.

TABLEAU 3.13 University Attendance by Gender

	(1) Men	(2) Women	(3) Men	(4) Women	(5) Men	(6) Women	(7) Men	(8) Women
Exp. ret. uncondi. empl.	0.019*	0.009	0.009	0.001	-0.008	-0.011	-0.007	-0.011
	(0.011)	(0.013)	(0.012)	(0.014)	(0.012)	(0.014)	(0.012)	(0.014)
Exp. max-min			0.048	0.004	0.027	-0.020	0.019	-0.055
			(0.042)	(0.067)	(0.043)	(0.067)	(0.043)	(0.069)
Actual wassce score					0.008***	0.005***	0.007***	0.004**
					(0.002)	(0.002)	(0.002)	(0.002)
Percvd prob. enter uni					0.000	0.002	0.000	0.002
					(0.002)	(0.003)	(0.002)	(0.003)
Percvd prob. finance uni					-0.000	0.001	-0.001	0.002
					(0.001)	(0.003)	(0.001)	(0.003)
Age							-0.033*	-0.061
							(0.019)	(0.042)
Father primary educ							-0.082	0.327
							(0.152)	(0.387)
Father JHS educ							-0.125	-0.018
							(0.102)	(0.237)
Father SHS educ							0.001	0.115
							(0.092)	(0.216)
Father post-sec educ							-0.083	0.122
							(0.116)	(0.211)
Mother primary educ							-0.022	0.046
							(0.133)	(0.221)
Mother JHS educ							0.010	-0.109
							(0.086)	(0.185)
Mother SHS educ							0.087	-0.024
							(0.099)	(0.166)
Mother post-sed educ							-0.050	-0.357*
							(0.115)	(0.187)
Number of siblings							0.009	0.002
							(0.013)	(0.028)
Born in rural area							-0.017	-0.097
							(0.057)	(0.099)
Constant	0.649***	0.591***	0.328	0.562	-0.074	0.176	0.667	1.528
	(0.079)	(0.107)	(0.299)	(0.482)	(0.337)	(0.520)	(0.542)	(1.105)
Region fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	342	150	342	150	342	150	342	150
R square	0.035	0.068	0.039	0.068	0.098	0.106	0.130	0.178
Adjusted R-squared	0.011	0.044	0.016	0.049	0.164	0.128	0.163	0.126
Average uni att.	0.637	0.62	0.637	0.62	0.637	0.62	0.637	0.62

Note : *** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at SHS level and are in parentheses. Anne Duplantier (2021)

TABLEAU 3.14 University Application by WASSCE Score

	(1) >aver.	(2) <aver.	(3) >aver.	(4) <aver.	(5) >aver.	(6) <aver.	(7) >aver.	(8) <aver.
Exp. ret. uncondi. empl.	0.008	0.013	-0.004	0.014	-0.019	0.004	-0.023	0.005
	(0.013)	(0.011)	(0.014)	(0.011)	(0.015)	(0.012)	(0.015)	(0.012)
Exp. max-min			0.056**	-0.003	0.050*	0.004	0.040	0.000
			(0.026)	(0.019)	(0.026)	(0.018)	(0.026)	(0.018)
Actual wassce score					0.011***	0.005***	0.010***	0.004***
					(0.002)	(0.001)	(0.002)	(0.001)
Percvd prob. admitted uni					0.002**	-0.000	0.002*	0.000
					(0.001)	(0.001)	(0.001)	(0.001)
Percvd prob. afford uni					0.001	0.002***	0.001	0.002***
					(0.001)	(0.001)	(0.001)	(0.001)
Age							-0.007	-0.019**
							(0.015)	(0.009)
Male							0.174***	0.044
							(0.054)	(0.031)
Father primary educ							0.035	-0.099
							(0.122)	(0.081)
Father JHS educ							-0.157**	-0.070
							(0.077)	(0.060)
Father SHS educ							-0.134*	-0.012
							(0.073)	(0.062)
Father post-sec educ							-0.044	-0.050
							(0.078)	(0.062)
Mother primary educ							0.106	-0.002
							(0.097)	(0.053)
Mother JHS educ							0.087	-0.003
							(0.067)	(0.047)
Mother SHS educ							0.102	-0.079
							(0.079)	(0.054)
Mother post-sec educ							0.122	-0.059
							(0.093)	(0.085)
Number of siblings							-0.007	-0.011**
							(0.010)	(0.005)
Born in rural area							-0.056	0.030
							(0.044)	(0.032)
Constant	0.462***	0.122	0.123	0.141	-0.878***	-0.211	-0.635	0.243
	(0.134)	(0.092)	(0.218)	(0.149)	(0.236)	(0.169)	(0.398)	(0.277)
Region fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	597	726	597	726	597	726	597	726
R-squared	0.061	0.044	0.068	0.045	0.123	0.074	0.156	0.090
Adjusted R-squared	0.041	0.027	0.047	0.026	0.099	0.052	0.124	0.060
Average uni appl.	0.556	0.220	0.556	0.220	0.556	0.220	0.556	0.220

Note : *** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at SHS level and are in parentheses. Anne Duplantier (2021)

application decision, while it is not related to the lower students' decision (Columns (3) and (4)). The difference in coefficients between the two groups is statistically significant at 10% level (p-value of the test equals to 0.094). For students with WASSCE scores superior to the average, an increase of 100 GHS in the gap between maximum and minimum expected returns is related to an increase of 5.6 units of the likelihood of application to university. However, after adding the other variables (Columns (5) to (8)), the perceived risk of earnings is not statistically significant and there is no difference in coefficients between the two groups.

The academic ability appears to be significantly and positively related to the decision to apply to university for both groups. However, the coefficient is around two times higher for respondents with a better score. When the WASSCE score increases by 1, the likelihood of application increases by 0.010 units for a student above the average, and 0.004 units for a student below the average. This difference in coefficients between the two groups is statistically significant (the p-value equals 0.0019). The perceived probability to be admitted into university appears to be significantly and positively related to the decision to apply for students with a score superior to the average only (Columns (5) and (6)). This difference between the two groups is statistically significant at 10% level. However, the magnitude is very close to zero (0.002 units). Moreover, the significance of this coefficient decreases after adding control variables (Column (7)). This suggests that the link between the perceived capacity to be admitted into university and the decision to apply is quite weak.

Finally, the perceived financial ability is not significantly related to the decision to apply for students with a higher score, whereas it matters for students with a lower WASSCE score. For these students, the relationship between the perceived probability to afford university is positively (although almost null) linked to their decision to apply. The difference in the coefficients between the two groups is statistically significant at 10% level. This can be explained by the fact that students with high WASSCE scores can attend university as non-fee-paying, which is not the case for students with low WASSCE scores. Students with less good WASSCE scores attend public university as fee-paying or attend private university. Therefore the capacity to pay for university influences the decision to apply. This financial constraint is confirmed by the significant and negative relationship between the number of siblings and the decision to apply to a university for students with low scores. Indeed, for individuals with a WASSCE score inferior to the average score, the more siblings they have, the less likely they are to apply to university.

Table 3.15 shows the results from the similar analysis for the attendance at university. The findings are quite similar to results for the application to university. For both groups, the academic ability is significantly and positively associated to the probability to attend university. However, the difference in coefficients is even higher than in Table 3.14 and still very significant. An increase in a

good student's WASSCE score of 1 increases their likelihood to attend university by 0.015 units. For the group of students with low WASSCE scores, the coefficient is almost null. This can be explained by the required score (superior to the average score) asked by the university to be admitted. For students whose WASSCE score is superior to the average, the probability to be admitted to university if the WASSCE score increases is higher than for students with a low score. Indeed, when the score of better students increases, these students have a higher probability to get a score required by the university to be admitted.

TABLEAU 3.15 University Attendance by WASSCE Score

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	>aver.	<aver.	>aver.	<aver.	>aver.	<aver.	>aver.	<aver.
Exp. ret. uncondi. empl.	0.012	0.008	-0.002	0.008	-0.014	0.001	-0.015	0.002
	(0.014)	(0.009)	(0.015)	(0.009)	(0.016)	(0.009)	(0.016)	(0.009)
Exp. max-min			0.033	-0.014	0.016	-0.016	0.021	-0.004
			(0.037)	(0.070)	(0.035)	(0.072)	(0.039)	(0.077)
Actual wassce score					0.016***	-0.003	0.015***	-0.003
					(0.003)	(0.003)	(0.003)	(0.003)
Percvd prob. enter uni					-0.001	0.003	-0.001	0.003
					(0.002)	(0.002)	(0.002)	(0.002)
Percvd prob. finance uni					-0.000	-0.003	-0.001	-0.002
					(0.002)	(0.002)	(0.002)	(0.002)
Age							-0.052**	-0.004
							(0.025)	(0.028)
Male							0.037	0.023
							(0.062)	(0.081)
Father primary educ							-0.068	0.034
							(0.160)	(0.233)
Father JHS educ							-0.136	-0.013
							(0.111)	(0.171)
Father SHS educ							0.043	-0.004
							(0.103)	(0.162)
Father post-sec educ							-0.070	0.098
							(0.105)	(0.177)
Mother primary educ							0.127	-0.003
							(0.142)	(0.168)
Mother JHS educ							0.142	-0.185
							(0.106)	(0.123)
Mother SHS educ							0.133	0.148
							(0.104)	(0.203)
Mother post-sed educ							-0.102	-0.032
							(0.105)	(0.264)
Number of siblings							-0.007	0.032
							(0.012)	(0.022)
Born in rural area							-0.067	-0.045
							(0.055)	(0.085)
Constant	0.680***	0.551***	0.456*	0.644	-0.597*	0.684	0.418	0.608
	(0.085)	(0.123)	(0.271)	(0.486)	(0.342)	(0.533)	(0.676)	(0.820)
Region fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	332	160	332	160	332	160	332	160
R square	0.020	0.100	0.022	0.100	0.119	0.114	0.192	0.172
Adjusted R square	0.016	0.026	0.017	0.020	0.075	0.015	0.117	0.005
Average uni att.	0.687	0.519	0.687	0.519	0.687	0.519	0.687	0.519

Note : *** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at SHS level and are in parentheses. Anne Duplantier (2021)

3.7.3 Sensitivity Analysis

Estimator Robustness

In order to check the sensitivity of the findings, I first replicate the main estimations by using probit and logit estimators. Table 3.16 presents the results of both probit and logit estimations, by gender, for the probabilities of university application and attendance. Results corroborate the main findings in terms of significance, sign and magnitude.

TABLEAU 3.16 Education Decision, Expectations about Labor Market and Ability - probit/logit

	Application to University				Attendance at University			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Probit	Probit	Logit	Logit	Probit	Probit	Logit	Logit
	Men	Women	Men	Women	Men	Women	Men	Women
Exp. ret. uncondi. empl.	-0.016	0.000	-0.016	0.003	-0.007	-0.011	-0.006	-0.009
	(0.012)	(0.014)	(0.012)	(0.014)	(0.012)	(0.014)	(0.012)	(0.014)
Exp. max-min	0.032	0.027	0.032	0.029	0.011	-0.009	0.013	-0.008
	(0.023)	(0.027)	(0.023)	(0.026)	(0.038)	(0.061)	(0.039)	(0.060)
Actual wassce score	0.010***	0.006***	0.010***	0.006***	0.007***	0.003*	0.007***	0.003*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
Percvd prob. enter uni	-0.000	0.001	-0.000	0.001	0.000	-0.000	0.000	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.003)	(0.002)	(0.003)
Percvd prob. finance uni	0.002***	0.002*	0.002**	0.002*	-0.001	0.001	-0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)
Age	-0.007	-0.024	-0.006	-0.026*	-0.030**	-0.050*	-0.030**	-0.050*
	(0.008)	(0.015)	(0.008)	(0.015)	(0.015)	(0.029)	(0.015)	(0.029)
Number of siblings	-0.004	-0.004	-0.005	-0.005	0.007	0.004	0.007	0.005
	(0.007)	(0.008)	(0.007)	(0.008)	(0.012)	(0.024)	(0.012)	(0.024)
Region fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	803	520	803	520	342	150	342	150

Note : *** p<0.01, ** p<0.05, * p<0.1. Average marginal effects are presented. Standard errors are clustered at SHS level and are in parentheses. Anne Duplantier (2021)

Sample Size Robustness

The second analysis of sensitivity comes from the variation of the sample size. Indeed, adding the variables of ability in the main estimations reduces the sample size from 3101 to 1275 observations. Indeed, these questions were asked to respondents only during 'In school' waves, before they graduated from SHS. Therefore, respondents from cohorts C_{2005}, C_{2006} and C_{2007} (that were never interviewed at SHS) do not have observations for these variables. Let's call this sub-sample of 1275 observations the 'ability's sub-sample'. As a second robustness check, I run the main estimations - with the expectations of the labor market - using the ability's sub-sample and compare it to the original results. The coefficient of the 'expected returns unconditional to employment' independent variable is relatively stable - 0.010 versus 0.013 - but is not significant. This is explained by the fact that the standard deviation rises when the sample size decreases - from 0.005 to 0.009. I then look to know whether the significance of expected returns disappears when adding ability's variables because of the effect of sample size or the effect of ability. I test the equality of expected returns

coefficients between the two estimations. The p-value equals to 0.019, suggesting that we reject the null hypothesis about equality of coefficients. In conclusion, the loss of significance of the expected returns comes from the ability effect and not the variation of the sample size. This can be explained by the fact that actual ability of student is the main determinant of their education decision.

TABLEAU 3.17 University Decision and Expectations about the Labor Market - Sample Size Robustness

	Application			Attendance		
	(1)	(2)	(3)	(4)	(5)	(6)
Exp. ret. uncondi. empl.	0.021**	0.014	0.010	0.011	0.009	0.006
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Exp. max-min		0.037*	0.027		0.031	0.018
		(0.019)	(0.018)		(0.033)	(0.033)
Age			-0.035***			-0.048***
			(0.009)			(0.018)
Male			0.153***			0.047
			(0.035)			(0.048)
Father primary educ			-0.070			-0.056
			(0.073)			(0.148)
Father JHS educ			-0.127**			-0.094
			(0.050)			(0.083)
Father SHS educ			-0.064			0.014
			(0.050)			(0.085)
Father post-sec educ			0.003			-0.006
			(0.048)			(0.093)
Mother primary educ			0.052			-0.003
			(0.050)			(0.116)
Mother JHS educ			0.044			0.010
			(0.043)			(0.076)
Mother SHS educ			0.023			0.098
			(0.052)			(0.086)
Mother post-sed educ			0.091			-0.120
			(0.068)			(0.101)
Number of siblings			-0.012**			0.001
			(0.005)			(0.011)
Born in rural area			-0.044			-0.046
			(0.031)			(0.046)
Constant	0.381***	0.138	0.854***	0.640***	0.433*	1.409***
	(0.061)	(0.143)	(0.218)	(0.070)	(0.241)	(0.447)
Region fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Cohort fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1323.000	1323.000	1323.000	492.000	492.000	492.000
R square	0.035	0.037	0.095	0.025	0.027	0.076
Adjusted R square	0.026	0.028	0.078	0.010	0.010	0.026
Average interest var. 0.300	0.300	0.300	0.607	0.607	0.607	

Note : *** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at SHS level and are in parentheses. Anne Duplantier (2021)

3.8 CONCLUSION

The decision to apply to and attend university after SHS is an important choice for SHS graduates that may have a strong impact on their life. Amongst the classical determinants of the education decision, one knows that the economic situation of the household, the education level of the parents, the number of siblings and the gender of the student play an important role. The recent economics literature considers the importance of the expectations in the individual's decisions. The emergence of the expectations allows us to capture the expected returns to education as a determinant of schooling decision.

In this chapter, I examined the role of expected returns to education on the decision to apply to and attend university in Ghana. As a novelty, I included the academic and perceived ability of the student as a determinant of the decision to go to university. This chapter also contributes to the literature by decomposing the process of the education decision in two steps : the application and the attendance. Information on the expectations in a developing country is quite sparse, in particular in the Sub-Saharan Africa where there is a strong lack of data. This research therefore contributes to the knowledge on expectations and education in a Sub-Saharan African country.

A linear model is used for estimating the link between decisions to apply to and attend university and the expected returns, as well as with both the perceived and academic ability. The results show that the expected returns to university are positively and significantly related to the application to university. But the relationship is no longer significant when adding the perceived and academic ability as other determinants. However, the ability has a strong relationship with the probability to apply to university. The results suggest that it is important to take ability into account when studying a schooling decision.

Regarding the probability to attend university, the results differ a little from the ones for the probability to apply. After controlling for the main individual characteristics, I found no significant link between the expected returns to university and the decision to attend university. After taking into account the ability, a significant link is found between the ability of the student and their decision to attend university. When looking at gender difference, the findings suggest that women are more constrained financially.

In conclusion, the findings suggest that the academic and perceived ability matters in the schooling decision. Moreover, as the results differ for the attendance and the application to university, I conclude that disaggregating the process of the education decision into two different steps is pertinent.

CHAPITRE 4
ATTRITION ANALYSIS

Résumé

Ce quatrième chapitre aborde la question technique du traitement de l'attrition qui touche la base de données en panel utilisée dans cette thèse. L'attrition et ses caractéristiques sont décrites et les déterminants de l'attrition sont analysés. Des tests de biais d'attrition sont mis en oeuvre et des corrections sont appliquées. Les résultats suggèrent que les individus sujets à l'attrition ont des caractéristiques différentes de ceux présents dans toutes les vagues. Ils sont en moyenne plus jeunes, ont une plus grande probabilité de venir de zones rurales et ont plus de frères et soeurs. De plus, ces personnes ont en général des parents moins éduqués et de moins bons résultats scolaires. Il est mis en évidence que les résultats d'intérêt du chapitre 3 souffrent d'un biais d'attrition. En effet, la probabilité de postuler et de fréquenter l'université est négativement liée à la probabilité d'être sujet à l'attrition. Ainsi une correction est appliquée aux résultats du Chapitre 3 et les résultats corrigés sont très proches de ceux non corrigés. Même si la base de données est sujette à l'attrition, les résultats restent pertinents et valides.

4.1 INTRODUCTION

A large number of researchers analyzing the labor market outcomes use a longitudinal database. This type of database allows tracing the dynamic of respondents' behaviors and controlling for unobserved fixed characteristics (Alderman et al., 2001). However, a key threat of panel analysis is attrition, which is the loss of sampled respondents dropping out of the survey. One consequence is that the sample size reduces over the time and the estimators lose precision (Watson, 2003). Winkel and Withers (2000) say that attrition is "the panel researcher's nightmare".

The problems occur when respondents leaving the survey are different from stayers and have characteristics that are linked to interest outcomes. In that case, attrition is likely to lead to a non-random selection sample and might bias the results, i.e. results are over- or underestimated. Therefore, checking attrition and the characteristics of attritors is crucial. Researchers should estimate the magnitude of the potential attrition bias in order to provide convincing and robust results based on panel data (Maluccio, 2004).

This dissertation uses the Ghana Opportunities for Transitioning Senior High School Students panel database which provides detailed information about education and labor market outcomes in Ghana. As in every panel database, the sample might be subject to selection because of attrition. In order to be sure that the results highlighted in this thesis are not biased due to attrition, this fourth chapter addresses the issue. First, attrition is described and characteristics of attritors are presented. Then, determinants of attrition are analyzed. Finally, I test for the presence of attrition bias and correct for potential bias using only the variables of Chapter 3.

Indeed, even if all three of the previous chapters would benefit from a description of attrition, results from Chapters 1 and 2 do not need to be tested for attrition bias. Chapter 1 addresses methodological issues and describes the characteristics of the sample. Therefore, there are no regression outcomes that need to be tested for attrition bias. However, descriptive statistics in Chapter 1 may be not representative due to selection bias from the first wave of survey. In order to check this potential bias, the magnitude and characteristics of attrition from the first wave are studied in Section 4 of this chapter. Moreover, the sub-sample used in Chapter 2 includes 1200 individuals interviewed in 2012. Amongst them, only 100 individuals left the survey and the previous wave of survey was used to replace the missing information. This reduces the likelihood of attrition bias.

This work contributes to the literature by drawing a sharp distinction between attrition types in a multi-waves panel database. Moreover, it analyzes attrition coming from unobservables. Indeed, it is often difficult to obtain a dataset including variables that affect attrition but are not related to the outcome of interest. In the literature, few panel databases have information that allows such analysis, such as variables characterizing the interview process. Fortunately, the database used in this

thesis includes such information. The results suggest that attritors have individual characteristics that differ from the stayers. An attrition bias is found in the results reported in Chapter 3, to which a correction is applied. The corrected results do not differ in general from the biased results, showing that we can trust the conclusion of Chapter 3.

In this chapter, I conduct an attrition analysis of the panel database collected and used in this thesis. In Section 2, the literature about attrition is discussed, followed by a description of the theoretical framework in Section 3. Then Section 4 presents the pattern of attrition in the database. The methodology is discussed in Section 5 before the results of the attrition analysis in Section 6. The correction of attrition is applied in Section 7 and Section 8 concludes the chapter.

4.2 LITERATURE

This section reviews the economic literature on attrition. The first part presents the main reasons for attrition suggested in the literature. The second and third parts review the papers analyzing attrition in developed countries and developing countries respectively. The last part presents what is done here and how it relates to the literature.

4.2.1 The Main Reasons of Attrition

According to Baigrie and Eyal (2014), fieldwork errors, non-contacts, refusals and death are the main reasons of attrition in a survey. In developed countries, the main reasons of attrition are refusals. Indeed, higher income levels produce a higher opportunity cost of time (Maluccio, 2004). In contrast, refusal rates are rather low in developing countries (Thomas et al., 2001). According to Baigrie and Eyal (2014), the major reason of attrition in developing countries is non-contact. Indeed, tracking respondents might be more difficult and costly than in developed countries as phone networks and postal records are less developed. Death of respondents also affect more surveys in developing countries because of higher mortality rates and lower life expectancy (Baigrie and Eyal, 2014). In these countries, flawed death records contribute to a higher likelihood of recording a migrant as dead. Finally, attrition in developing countries is influenced by a range of factors such as cultural aspects (Maluccio, 2004) or the match between enumerator and respondents on some aspects such as gender and race.

4.2.2 Attrition in Developed Countries

Attrition has gathered much interest in the late 1990s and has been the focus of specific literature. At first, research focused on attrition in panel surveys in developed countries. Lillard and Panis (1998), Zabel (1998) and Fitzgerald et al. (1998) use a panel database called the Michigan Panel

Study of Income Dynamics (PSID) from households in the United States of America (USA) with waves from 1968 to 1989. Winkel and Withers (2000) and Watson (2003) focused respectively on the Netherlands and 14 of the European Union countries. In general, attrition rates are large : from 28% to 51% in the USA and from 11% to 42% in Europe.

First, the literature analyzes the characteristics of attritors. In general, authors find significant differences between attritors and stayers, suggesting that attrition is mainly non-random and selective. In the USA, the database PSID shows that attrition is larger for individuals of lower socio-economic status and is related to a low level of education (Zabel, 1998; Lillard and Panis, 1998). This contradicts the assertion of Maluccio (2004) previously presented. But according to Watson (2003), differences by economic status and education are rather small in Europe. Attrition is larger for young adults, not married and living alone (Watson, 2003). Lillard and Panis (1998) find that married respondents have higher continuation rates in the PSID database, but older people have a higher probability to drop out.

Second, literature on attrition searches for attrition bias and examines the determinants of attrition. Fitzgerald et al. (1998) propose a test for attrition bias that estimates attrition using auxiliary variables. Using this methodology, Lillard and Panis (1998) find that results are slightly affected by attrition bias. Estimating determinants of attrition with a probit, Zabel (1998) finds that demographic variables have little impact on attrition in the USA, but the characteristics of interviewers and the interview process have significant impacts. They conclude that there is a little attrition bias on labor market behavior. Similarly, Fitzgerald et al. (1998) show that attrition does not seriously affect the representativeness of the sample of the PSID database. And according to Lillard and Panis (1998), the estimations of household income dynamics, marriage formation and dissolution, as well as adult mortality risk, are very slightly affected by attrition bias.

4.2.3 Attrition in Developing Countries

In the early 2000s, economists started studying attrition in databases from developing countries. Falaris (2003) examines the pattern of attrition from three Living Standards Measurement Study panel databases from Peru, Ivory Coast and Vietnam. In these countries, attrition rates vary from 15% in Vietnam to 30% in Peru. The comparison method is used in order to evaluate the impact of attrition on results. This method involves testing the equality of the regression coefficients for stayers and attritors. In general, Falaris (2003) finds that attrition is patterned, as individuals with less schooling and a lower labor force participation rate are more likely to attrit. But attrition does not have a major impact on estimates of schooling attainment, labor force participation and fertility in these three countries.

The majority of the literature on attrition in developing countries finds similar results (Alderman

et al., 2001; Baigrie and Eyal, 2014; Baulch and Quisumbing, 2011; Outes-Leon and Dercon, 2009; Magruder and Nattrass, 2006; Fuwa, 2011). It appears that attrition is non-random and strongly varies according to the countries (5% in Bangladesh versus 41% in South Africa and Kenya). Studying the differences between attritors and non-attritors, authors find a higher probability to attrit for younger Kenyans with higher schooling. In South Africa, attritors are older, more educated, richer and more likely to be male. In Madagascar, stayers are older, more likely to be married and more likely to be the head of a household.

Alderman et al. (2001) estimate the determinants of attrition with a probit and find that family background and health outcome variables (such as child development, fertility and a family planning program) do not predict attrition. According to Baulch and Quisumbing (2011), baseline variables such as the age of household head and household demographic variables explain about 13% of the panel attrition in Bangladesh. In South Africa, Magruder and Nattrass (2006) show that the impact of both individual and household characteristics on the probability of attrition is relatively small. They find that the probability to attrit is higher for younger men and women living in smaller and poorer households. With another survey from South Africa, Maluccio (2004) finds that few community characteristics are strong predictors of attrition.

Finally, the majority of these articles finds very limited evidence of attrition bias on the estimates. Outes-Leon and Dercon (2009) use an attrition probit test [1] and a test analysing the bias that attrition might cause on the model coefficients. [2] According to the authors, attrition is non-random as attriting households tend to be poorer and less educated, but it does not bias the results of child anthropometric and school enrollment models. In Bolivia, Kenya and South Africa, the association between family background and child development outcome variables is not affected significantly by attrition (Alderman et al., 2001). Lee (2003) identifies this as the “neutrality of attrition". Similar results on labor market analysis in South Africa are found by Magruder and Nattrass (2006). Fuwa (2011) assesses that the determinants of per-capita household consumption in Philippines and its growth rate can be estimated based on the stayer-only sample without serious bias.

Baigrie and Eyal (2014) and Maluccio (2004) are the only ones who mitigate the evidence that results are not biased by attrition. Studying both labor market and health outcomes in South Africa, Baigrie and Eyal (2014) find a moderate evidence of attrition bias in these specifications. Still in South Africa but with another database, Maluccio (2004) explains that attrition leads to statistical bias in household-level expenditure results.

1. This test estimates the likelihood of attrition with a probit and is presented in more detail in the methodological section.

2. This test estimates the interest outcomes separately for the full sample and the non-attriting sample, and tests for the equality of the coefficients individually and jointly. It is presented in more detail in the methodological section.

4.2.4 Position in the Literature

Following the literature on attrition, the characteristics of attritors are described using the usual methods. This chapter differs from the literature by distinctly defining the attrition types. Indeed, four types of respondents are defined instead of two. Usually, studies only distinguish stayers from attritors. One reason for this is that the panel database used has only two waves of survey. Another is that even with panel databases including several waves of survey, few researchers continue to track respondents that attrit once. In our survey, however, the choice has been made to follow respondents that attrit for a wave to be able to survey them the next wave. As a consequence, we are able to distinguish four types of respondents according to their attrition behavior. These types are presented in more details in Section 4. Moreover, instead of only studying attrition coming from observables as in most of the literature, I am able to analyze attrition coming from unobservables. Indeed, our database includes variables which describe the survey process. According to Maluccio (2004), such variables can be considered as auxiliary variables that are required for testing attrition coming from unobservables.

4.3 THEORETICAL FRAMEWORK

This section presents the theoretical foundations of the attrition analysis.

Perfect random samples are rather scarce. The way in which surveys are implemented, as well as the behavior of respondents, often result in a non-random sample. As explained in Wooldridge (2002), there are several mechanisms leading to sample selection. The first one occurs when researchers cannot reach the whole sample during the first wave of a panel survey. Even in the case of a randomized sample, the loss of respondents might be non-random. In the survey used in this dissertation, there is a risk of such sample selection. Indeed, the sampling is at Senior High School (SHS) level, and interviews were conducted at school during school hours. Therefore, the absence of respondents could be due to non-random characteristics. For example, a student from a rural household might miss school during the harvest period in order to help their family in the field. This might also occur if a student from a poor household is required to work occasionally in order to contribute to the family income. The group of respondents who were present at school the day of the interview constitutes a sub-sample of the initial randomized sample.

Another mechanism of sample selection is attrition (Vaillant, 2013). In a panel data, each wave generally suffers the loss of observation units. Fitzgerald et al. (1998) are the first to propose a model designing attrition. They consider a panel study of T waves, surveying the same sample of individuals every year noted $t = 1, ..., T$. An attritor of the wave t is a respondent who was

interviewed in every wave $t = 1, ..., t-1$ but was not interviewed from the year t onward.[3] Attrition of this respondent is denoted A_t.

Consider an interest outcome y depending on variables x included in the vector X. The model of interest in time t is :

$$y_t = \beta_0 + \beta_1 X_t + \epsilon_t \tag{4.1}$$

But y_t is observed only if $A_t = 0$, i.e. if respondent is not an attritor in the wave t. Because of attrition, one can estimate Equation 4.1 only for respondents who are interviewed during the wave t. In the case where the value of y_t is missing because of attrition, we have $A_t = 1$.

The probability of attrition is defined as :

$$Pr(A_t = 1 | y_t, X_t, Z_t) \tag{4.2}$$

where Z_t is a set of auxiliary variables that are not included in X_t. According to Alderman et al. (2001), Z_t can include lagged values of the dependent variable, fixed characteristics of the respondent, lagged time-varying characteristics and variables that do not require the completion of an interview, such as interviewer characteristics and location of residence.

If we linearize the probability of attrition, we get :

$$A_t^* = \delta_0 + \delta_1 y_t + \delta_2 X_t + \delta_3 Z_t + \nu_t \tag{4.3}$$

In practice, the probability of attrition A^* is not measurable - it is a latent variable - and we only observe whether a respondent has been interviewed or not. So we have :

$$\begin{aligned} A_t &= 1, && \text{if } A_t^* \geq 0 \\ &= 0, && \text{if } A_t^* < 0 \end{aligned} \tag{4.4}$$

This attrition selection model determines the likelihood of an observation being observable in Equation 4.1. Attrition is totally random if attrition is not related to y_t, X_t and Z_t. On the contrary, there is an attrition bias in Equation 4.1 if attrition depends on one of these variables or the error terms from Equations 4.1 and 4.3, ϵ and ν, are correlated. This occurs when a variable other than x affects both attrition and the interest variable of Equation 4.1. This specification explains why attrition bias is model specific and non sample specific. Indeed, the fact that ϵ is correlated with ν depends on the model estimated. The other implication is that non-random attrition does not necessarily lead to biased results.

3. I present only the simplest case of attrition, but there are some others cases of attrition in the database. See the next section.

Fitzgerald et al. (1998) distinguish two types of factors conducting to attrition : observable and unobservable characteristics. There is attrition on observables if variables in Z are not independent of $\epsilon|X$ and ν are independent of $\epsilon|X$. It means that all variables that might affect both the attrition equation (Equation 4.3) and the main equation (Equation 4.1) are observables for both attritors and stayers. They can be time-invariant variables, as well as time-variant characteristics measured before attrition and included as lagged variables. We can rewrite it as follows :

$$Pr(A_t = 0|y_t, X_t, Z_t) = Pr(A_t = 0|X_t, Z_t) \tag{4.5}$$

and Equation 4.3 becomes :

$$A_t^* = \alpha_0 + \alpha_1 X_t + \alpha_2 Z_t + \mu_t \tag{4.6}$$

On the other hand, attrition on unobservables occurs in the opposite situation, when Equation 4.5 does not hold. In this case, variables in Z are independent of $\epsilon|X$ but ν is not independent of $\epsilon|X$. It means that the probability of being an attritor is correlated to y_t and unobserved factors included in ϵ_t. This occurs if unobserved characteristics such as individual motivation or risk aversion affect both attrition and the variable of interest. It also occurs if attrition is related to some characteristics that changed between waves $t-1$ and t and are not measured for attritors.

4.4 DESCRIPTION OF ATTRITION

This section describes how the database used in this thesis is affected by attrition. In the first part, the sample is decomposed according to the number of interviews done. The second part presents the extent of attrition in the database. The third part decomposes attrition to understand the reasons of attrition. The last part presents the probability of being interviewed.

4.4.1 Composition of the Sample

First, the sample's composition is described by focusing on the number of times respondents have been interviewed. Table 4.1 presents the structure of the database for the full sample and then by cohort. Panel A of Table 4.1 shows the frequency and the percentage of respondents that have been never interviewed or interviewed one to five times. Panel B of Table 4.1 presents the same information for the theoretical sample, i.e. without attrition. The sample should be composed of 7616 individuals forming three groups of respondents. The respondents from cohorts C_{2005} and C_{2006} should have been interviewed three times. Those from cohorts C_{2007}, C_{2008} and C_{2011} should have been interviewed four times. The respondents from cohorts C_{2009} and C_{2010} should have been interviewed five times.

TABLEAU 4.1 Sample Structure

Waves	All	C_{2005}	C_{2006}	C_{2007}	C_{2008}	C_{2009}	C_{2010}	C_{2011}
Panel A : Actual Sample								
No one	1048	418	315	113	53	61	63	25
	[13.76]	[38.42]	[28.95]	[10.39]	[4.87]	[5.61]	[5.76]	[2.3]
1 wave	528	37	16	42	81	66	101	185
	[6.93]	[3.4]	[1.47]	[3.86]	[7.44]	[6.07]	[9.28]	[17]
2 waves	719	125	159	47	69	70	68	181
	[9.44]	[11.49]	[14.61]	[4.32]	[6.34]	[6.43]	[6.25]	[16.64]
3 waves	2060	508	598	184	247	156	151	216
	[27.05]	[46.69]	[54.96]	[16.91]	[22.7]	[14.34]	[13.88]	[19.85]
4 waves	2579	-	-	702	638	390	368	481
	[33.86]	-	-	[64.52]	[58.64]	[35.85]	[33.82]	[44.21]
5 waves	682	-	-	-	-	345	337	-
	[8.95]	-	-	-	-	[31.71]	[30.97]	-
Panel B : Theoretical Sample								
3 waves	2176	1088	1088	-	-	-	-	-
	[28.57]	[100]	[100]	-	-	-	-	-
4 waves	3264	-	-	1088	1088	-	-	1088
	[42.86]	-	-	[100]	[100]	-	-	[100]
5 waves	2176	-	-	-	-	1088	1088	-
	[28.57]	-	-	-	-	[100]	[100]	-

Note : One observation per respondent ($N = 7616$). Numbers presented are frequencies. Percentage are in brackets and are computed by column (for the full sample and by cohort). C_{20XX} refers to the cohort of students that started SHS in 20XX. Anne Duplantier (2021)

In reality, the sample contains six groups of respondents interviewed between zero and five times. We count 1048 respondents who were never interviewed, which represents 13.7% of the sample. Cohorts C_{2005} and C_{2006} have the highest rates of no interview, 38.4% and 28.9% respectively. One explanation lies in the fact that these students graduated from SHS two or three years before the first round of data collection. As the starting information, such as names and addresses, came from schools' listings it was more difficult to track these respondents. Indeed, it is possible that the information was now out of date, or that the respondents had moved after SHS. For cohorts C_{2008} to C_{2011} that were still at school during the first interview, the rate of no interview is lower, varying between 2.3% to 5.7%.

Table 4.2 shows the number of respondents with missing waves. Around 47% of the sample has been interviewed the right number of times (the 'No missing' column). However, one wave of interview is missing for 22.1% of the sample, and 15.5% misses three waves.

TABLEAU 4.2 Number of Missing Waves

	No missing	1 missing	2 missing	3 missing	4 missing	5 missing
Respondents	3609	1689	657	1179	358	124
Percentage	(47.39)	(22.18)	(8.63)	(15.48)	(4.7)	(1.63)

Note : One observation per respondent ($N = 7616$). Standard deviations are in parentheses. Anne Duplantier (2021) Anne Duplantier (2021)

Table 4.3 summarizes respondents by theoretical number of waves and actual ones. There are 733 respondents that have never been interviewed who should have been interviewed three times. This group of respondents represents 9.6% of the total sample, and accounts for 69.9% of the respondents who have never been interviewed. In the sample, 23.9% of the respondents have been interviewed four times and should have been interviewed four times.

4.4.2 Attrition in the Sample

Table 4.4 gives an overview of the extent of attrition. Panel A shows the retention rates of the whole panel. The retention rates is the exact inverse of attrition rate, i.e. the rate of respondents that have been successfully interviewed. The panel sample is composed of 21 872 respondents instead of 30 464, suggesting that 71.8% of the respondents were successfully interviewed. There are strong differences across cohorts. The retention rate is only 55.5% for cohort C_{2005}, while 80.7% of cohort C_{2008} was reached.

In order to check whether there are differences in retention by place of interview (at school versus outside of school), Panel B presents retention rates decomposed by type of questionnaire ('In School' versus 'Go Transition'). One could expect to obtain higher retention at school than outside.

TABLEAU 4.3 Actual and Theoretical Number of Waves

Theoretical	Actual Number					
Number	No one	1 wave	2 waves	3 waves	4 waves	5 waves
3 waves	733	53	284	1106	-	-
	[69.94]	[10.04]	[39.5]	[53.69]	-	-
	(9.62)	(0.69)	(3.73)	(14.52)	-	-
4 waves	191	308	297	647	1821	-
	[18.23]	[58.33]	[41.31]	[31.41]	[70.61]	-
	(2.51)	(4.04)	(3.9)	(8.49)	(23.91)	-
5 waves	124	167	138	307	758	682
	[11.83]	[31.63]	[19.19]	[14.9]	[29.39]	[100]
	(1.63)	(2.19)	(1.81)	(4.03)	(9.95)	(8.95)
Total	1048	528	719	2060	2579	682

Note : One observation per respondent ($N = 7616$). Numbers presented are frequencies. Percentage with respect to the column are in brackets. Percentage with respect to the full sample are in parentheses. Anne Duplantier (2021)

TABLEAU 4.4 Retention Rate

	All	C_{2005}	C_{2006}	C_{2007}	C_{2008}	C_{2009}	C_{2010}	C_{2011}
Panel A : No Decomposition								
Full Panel	21872	1811	2128	3496	3512	3959	3847	3119
($N^* = 30464$)	[71.8]	[55.48]	[65.2]	[80.33]	[80.7]	[72.78]	[70.72]	[71.67]
Panel B : By Type of Questionnaire								
In School	6964	-	-	-	912	2216	2193	1643
($N^* = 9792$)	[71.12]	-	-	-	[83.82]	[67.89]	[67.19]	[75.51]
Go Transition	14908	1811	2128	3496	2600	1743	1654	1476
($N^* = 20672$)	[72.12]	[55.48]	[65.20]	[80.33]	[79.66]	[80.10]	[76.01]	[67.83]
Panel C : By Wave of Survey								
In School 2011	2723	-	-	-	912	906	905	-
($N^* = 3264$)	[83.43]	-	-	-	[83.82]	[83.27]	[83.18]	-
Go Transition 2011	2224	629	719	876	-	-	-	-
($N^* = 3264$)	[68.14]	[57.81]	[66.08]	[80.51]	-	-	-	-
In School 2012	2534	-	-	-	-	779	765	990
($N^* = 3264$)	[77.63]	-	-	-	-	[71.60]	[70.31]	[90.99]
Go Transition 2012	3087	610	724	902	851	-	-	-
($N^* = 4352$)	[70.93]	[56.07]	[66.54]	[82.9]	[78.22]	-	-	-
In School 2013	1707	-	-	-	-	531	523	653
($N^* = 3264$)	[52.30]	-	-	-	-	[48.81]	[48.07]	[60.02]
Go Transition 2014	5337	572	685	844	862	864	834	676
($N^* = 7616$)	[70.08]	[52.57]	[62.96]	[77.57]	[79.23]	[79.41]	[76.65]	[62.13]
Go Transition 2018	4260	-	-	874	887	879	820	800
($N^* = 5440$)	[78.31]	-	-	[80.33]	[81.53]	[80.79]	[75.37]	[73.53]

Note : All observations for each respondent included. Numbers presented are frequencies. Percentage are in brackets and represents retention rate (i.e. non-attrition rate). N^* is the theoretical sample size. In the 'Go Transition 2018' wave, C_{2005} and C_{2006} cohorts were not included in the sampling frame as retention rates in previous waves were low. C_{20XX} refers to the cohort of students that started SHS in 20XX. Anne Duplantier (2021)

Indeed, if the attendance rate is high, the likelihood to find respondents at school could be higher than the probability to find respondents outside of school. We find in the data that retention is 71.1% among respondents interviewed at school versus 72.1% for the group of respondents interviewed outside of school. However, there are differences by cohort. The retention rate of the C_{2008} and C_{2011} cohorts is highest at school whereas it is the opposite for C_{2009} and C_{2010}.

In order to understand better the differences in retention between respondents at school and the ones outside, Panel C presents retention rates per year and place of interview. Retention rates in 2013 are particularly low. Indeed, for financial reasons, three regions were not interviewed this year : Northern, Upper West and Upper East. In the seven other regions, retention rates vary between 56% (in Ashanti) and 76% (in Volta). Moreover, retention rates of the 2018 wave are higher by construction. Indeed, the C_{2005} and C_{2006} cohorts were not included in the sampling frame as retention rates in previous waves were low and decreasing. The difficulty in reaching these two first cohorts (C_{2005} and C_{2006}) can be explained by the fact that they were not interviewed at school for the first wave of the survey, unlike the other cohorts. Moreover, students from C_{2005} and C_{2006} had graduated from SHS two or three year prior the first wave of survey. As sampling was built from SHS, the starting information, such as names and addresses, came from schools' listings. Therefore, information about these respondents is more likely to be out of date, making it more difficult to survey respondents from these cohorts.

From Panel C, we can compare retention rate by years between the two types of interview. In years where interviews were conducted both at school and outside (in 2011 and 2012), retention rates are on average higher at school than outside. This is consistent with what was expected. Moreover, we can look at the evolution of the retention rate by cohort and type of interview across time. For each cohort interviewed at least twice at school (C_{2009}, C_{2010} and C_{2011}), the first year of interview at school has a higher retention rate than the second year. Therefore, even at school, there is attrition. One explanation is that students change schools. Another is that students are older and attend school less frequently as they work occasionally.

Figure 4.1 reports the reasons for non-interview. Two categories can be distinguished : exogenous and endogenous reasons. The endogenous reasons are refusals and non-contacts. They are likely to be related to aspects of the survey process such as surveying method, sampling strategy and follow-up rules. Among the non-interview observations, 1.5% are missing because respondents refused to participate in the survey. Moreover, inability to contact respondents represents 95% of non-interview cases, despite all the efforts made to track the respondents.

Exogenous reasons arise from the natural dynamic of the population, such as the death and the migration outside of the country. These events are *a priori* not influenced by the survey. Out of the 8592 non-interview, 42 observations are missing because the respondent died and 217 because

the respondent migrated abroad. They represent respectively 0.5% and 2.5% of the non-interview observations. Information was collected from the respondent's acquaintances for whom we have contact details from previous waves. However, it is very likely that other respondents may have died or migrated without our knowledge. Indeed, if we have no contact information or if the contact's details are now out of date, we will have no access to information concerning the respondents. Some respondents are probably classed as 'unable to contact' whereas they actually died or migrated. The statistics of dead and migrant respondents is therefore likely to be underestimated and the non-contact category overestimated.

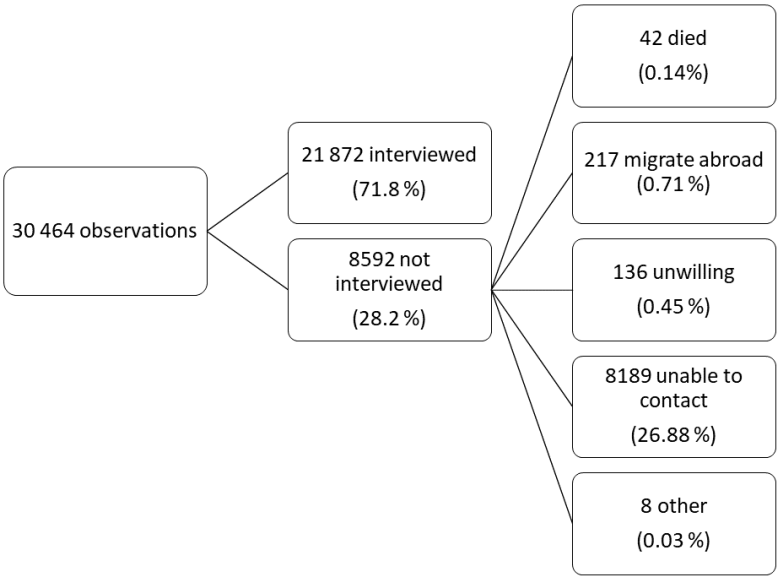
4.4.3 Decomposition of Attrition

Kalton and Brick (2000) consider two types of attrition. Attrition is permanent when a respondent leaves definitively the survey. Attrition is considered as temporary when a respondent leaves and re-enters the survey at different waves. This leads to four categories of respondents, depending on the frequency of presence of a respondent in the survey. The 'permanent stayer' is successfully interviewed in each wave, i.e. there is no attrition. The 'temporary stayer' is interviewed in certain waves but not in each of them - they temporarily leave the survey then re-enter it. In this case, attrition is temporary. The 'wave attritor' leaves the survey definitively after being interviewed in the first or subsequent waves : attrition is permanent for this type of individual. The 'total attritor' has never been interviewed even though they were included in the sample : attrition is permanent in this case.

Table 4.5 presents the frequency and percentage of each category of respondent. Column (1) reports that 47.3% of the sample has been interviewed every time they were due to be - permanent stayers - while 13.7% of the sample has never been interviewed - total attritors. The wave-attritors, i.e. respondents who dropped out of the survey having been interviewed, represent 11.7% of the respondents. Finally, 27% of the sample is composed of temporary stayers, i.e. individuals who enter and leave the survey, depending the waves.

The remaining columns shed light on differences between cohorts. Cohort C_{2010} has the lowest rate of permanent stayers with 31%, while C_{2007} has the highest with 64.5%. C_{2011} is simultaneously the cohort with the smallest total attritors rate and the highest rate of wave-attritors. Cohorts C_{2009}, C_{2010} and C_{2011} have the highest rate of temporary stayers (from 30% to 54%) compared to the four other cohorts (from 6% to 24%). Table C.1 in the Appendix presents the types of respondent in more detail, according to the theoretical number of waves.

FIGURE 4.1 Reasons for Non-Interview



Anne Duplantier (2021)

TABLEAU 4.5 Categories of Respondent

	All	C_{2005}	C_{2006}	C_{2007}	C_{2008}	C_{2009}	C_{2010}	C_{2011}
Permanent Stayer	3 609	508	598	702	638	345	337	481
	[47.39]	[46.69]	[55.01]	[64.52]	[58.69]	[31.71]	[31]	[44.33]
Wave Attritor	898	94	83	89	133	99	148	252
	[11.79]	[8.64]	[7.63]	[8.18]	[12.23]	[9.1]	[13.61]	[23.22]
Temporary Stayer	2061	68	92	184	264	583	540	330
	[27.06]	[6.25]	[8.46]	[16.91]	[24.29]	[53.58]	[49.68]	[30.41]
Total Attritor	1048	418	315	113	53	61	63	25
	[13.76]	[38.42]	[28.98]	[10.38]	[4.87]	[5.61]	[5.79]	[2.3]

Note : One observation per respondent ($N = 7616$). Percentage are in brackets and computed by columns. A permanent stayer was interviewed every time she should be. A wave attritor was interviewed and then dropped out of the panel. A temporary stayer left the panel temporarily and re-entered. A total attritor was never interviewed. C_{20XX} refers to the cohort of students that started SHS in 20XX. Anne Duplantier (2021)

4.4.4 Probability of Being Interviewed

Table 4.6 shows the probability to be interviewed according to the number of the wave. The likelihood of being interviewed during the theoretical first wave is 77.9%. This probability decreases with the rank of the wave. Indeed, 69.3% of respondents have been interviewed during their theoretical second wave of interview, and the probability is only 61.6% for the third wave. It shows the challenges of longitudinal surveys to be able to reach respondents over time. This probability goes up to 78% for the fourth and fifth wave. One explanation for this is that the first two cohorts have been removed from the sample in 2018 as they were too difficult to reach (as explained above). Another explanation is that there is a threshold at which the probability of being interviewed rises.

TABLEAU 4.6 Likelihood of Interview by Wave

Wave	First	Second	Third	Fourth	Fifth
Likelihood of Interview	77.95	69.38	61.62	78.29	78.08

Note : One observation per respondent ($N = 7616$). Probabilities are percentage. Anne Duplantier (2021)

Table 4.7 presents the sequences of interview for the full sample and by cohort. The likelihood to be consecutively interviewed after being interviewed is 45%. Moreover, 14% of respondents are consecutively not interviewed after not being interviewed. It suggests that there is a virtuous circle of being interviewed and a vicious circle of not. However, the probability to be successively interviewed after not being interviewed is 7.1%.

TABLEAU 4.7 Probabilities of Sequences of Interview

	All	C_{2005}	C_{2006}	C_{2007}	C_{2008}	C_{2009}	C_{2010}	C_{2011}
Successfully interviewed after being successfully interviewed	45.19	34.13	40.35	55.91	53.54	44.54	43.27	41.27
Successfully interviewed after not being successfully interviewed	7.12	2.08	2.82	4.30	6.20	11.58	10.81	7.65
Not successfully interviewed after being successfully interviewed	8.50	3.83	3.86	4.34	6.78	12.08	12.37	12.02
Not successfully interviewed after not being successfully interviewed	14.20	26.62	19.64	10.45	8.48	11.80	13.55	14.06
Successfully interviewed during the first wave	19.19	19.27	22.03	20.13	20.96	16.65	16.64	22.75
Not successfully interviewed during the first wave	5.51	14.06	11.31	4.87	4.04	3.35	3.36	2.25
Total	100	100	100	100	100	100	100	100

Note : All observations for each respondent included ($N = 30464$). Probabilities are percentage. C_{20XX} refers to the cohort of students that started SHS in 20XX. Anne Duplantier (2021)

Table 4.8 summarizes some statistics on attrition by SHS region. Panel A presents the retention rate by region. Greater Accra, which is the richest and most economically attractive region, has the highest retention rate at 76.64%. Upper West, one of the poorest and most rural regions, has the

lowest retention rate at 60.63%. Retention rates at school are the lowest in the Northern, Upper East and Upper West regions, at between 58.3% and 61.3%. These regions are the most rural and the poorest of the country. It is likely that attendance is less systematic than in other regions. Consistently, the retention rate at school is the highest in Greater Accra at 80.3%. Another explanation for the regional differences in retention rate is that the surveyor team was based in Accra. For in-person surveys (in 2011, 2012 and 2013), this may have a consequence on retention rates. In regions far from Accra (that are also the poorest), the surveyors were less likely to have been managed by supervisors based in Accra. However, variation in retention rates outside of school between regions is lower than for retention rates in school, suggesting that the main reason for the regional difference is the differing attendance rates.

TABLEAU 4.8 Retention Rate and Category of Respondent by SHS Region

	Ashanti	Brong Ahafo	Central	East.	Greater Accra	North.	Upper East	Upper West	Volta	West.
Panel A : Retention Rate ($N_F^* = 30464$, $N_I^* = 9792$ and $N_G^* = 20672$)										
Full Panel	71.13	67.68	72.56	75.16	76.64	68.85	65	60.63	74.69	71.75
In School	68.23	64.54	75.56	77.57	80.32	58.33	61.39	59.44	73.89	73.74
Go Transition	72.51	69.17	71.14	74.01	74.89	73.83	66.71	61.18	75.07	70.81
Panel B : Category of Respondent ($N^* = 7616$)										
Perm. Stayer	46.13	45.36	50.48	55.27	52.53	36.31	24.29	24.29	53.04	48.54
Wave Attritor	9.15	11.67	11.31	11.52	9.52	11.71	13.21	17.86	12.59	16.56
Temp. Stayer	29.39	26.19	21.90	20.45	25.45	42.66	51.79	41.43	22.05	22.56
Total Attritor	15.33	16.79	16.31	12.77	12.50	9.33	10.71	16.43	12.32	12.34

Note : Probabilities are percentage. N^* is the theoretical sample size, N_F^* is the theoretical sample size of the full sample, N_I^* is the theoretical sample size of the 'In School' sample, and N_G^* is the theoretical sample size of the 'Go Transition' sample. Anne Duplantier (2021)

Panel B of Table 4.8 presents categories of respondent by region. For permanent stayers and temporary stayers, there are clearly two regional groupings. On one hand, Ashanti, Brong Ahafo, Central, Eastern, Greater Accra, Volta and Western have a proportion of permanent stayers between 45% and 55% and a rate of temporary stayers between 20% and 29%. On the other, Northern, Upper East and Upper West have the lowest rates of permanent stayers : between 24% and 36%, and the highest rates of temporary stayers : between 41% and 52%. The proportion of wave-attritors is very similar for each region - between 9% and 13% - except in Western and Upper West : 16.5% and 17.8%. The rates of total attritors is the lowest in Northern - 9.3% - and the highest in Brong Ahafo - 16.7%.

Finally, Table 4.9 provides statistics on attrition by gender. Panel A summarizes retention rates. Interestingly, these are higher for girls than boys : 73.9% versus 71.8% (difference statistically significant at 1% level). This gender difference stays statistically significant for retention rate at school. However, there is no statistically significant difference in retention rate outside school between girls and boys. Panel B focuses on category of respondent. There are more permanent stayers in the female group than in the male group : 50.3% versus 46.4% (difference statistically significant). Males

have a higher probability to leave either temporary or permanently the panel (difference statistically significant) : their presence in the panel is more unstable.

TABLEAU 4.9 Retention Rate and Category of Respondent by Gender

	(1) Female	(2) Male	(3) P-value
Panel A :	Retention Rate		
Full Panel ($N^* = 30464$)	73.9	71.81	0.000
In School ($N^* = 9792$)	74.61	69.81	0.000
Go Transition ($N^* = 20672$)	73.55	72.74	0.207
Panel B :	Category of Respondent		
Permanent Stayer	50.35	46.45	0.000
Wave Attritor	12.76	11.30	0.028
Temporary Stayer	25.08	28.86	0.000
Total Attritor	11.81	13.39	0.023
Observations ($N^* = 7616$)	2 971	4 549	

Note : Probabilities are percentage. N^* is the theoretical sample size. Over the 7616 respondents, there are 96 individuals for which the gender is unknown (1.3% of the sample). Column (3) is the p-value of the t-test difference between Columns (1) and (2). Anne Duplantier (2021)

4.5 METHODOLOGY FOR ANALYZING ATTRITION

This section presents the methodology employed to analyze attrition. The first part focuses on the characteristics of attritors. The second part explains how to test the pattern of attrition. The last part presents the method used to correct the results affected by attrition bias.

4.5.1 Characteristics of Attritors

Studying the main characteristics of the respondents who dropped out of the sample is crucial for understanding the mechanism of attrition. Attrition analysis usually starts with a presentation of the differences in characteristics of respondents. Following the literature, I present the statistics of characteristics for the whole sample, then by sub-group of attritors and non-attritors. Two methods are used for testing the difference between the two sub-groups. First, following what is done traditionally, I run a t-test for each variable and a chi-square test for the categorical variables. Second, in order to control for the cohort effects, the likelihood of being an attritor depending on each characteristics is separately estimated by (Ordinary Least Square) OLS with cohort fixed effects.

4.5.2 Pattern of Attrition

Literature on attrition distinguishes two major types of procedure to test the pattern. The first test is called the FGM test (because of the authors Fitzgerald, Gottschalk and Moffitt). For this method, Fitzgerald et al. (1998) suggest estimating a probit of the attrition equation (Equation 4.3). The dependent variable is equal to 1 if the respondent is an attritor for the wave t and 0 otherwise. The independent variables are the baseline values of X_t (i.e. all variables that can impact y_t from Equation 4.1). Zabel (1998) uses the same method and suggests the inclusion in X_t of lagged variables and the first-period values of the time-varying variables, as information for the current period is not available for attritors. Thus the results are conditional on the first period characteristics (Zabel, 1998).

In order to estimate the attrition equation, one must finding suited variables to proxy Z_t as defined in the theoretical section. For attrition coming from observables, Z_t has to affect both A_t and y_t conditionally to X_t. In other words, Z_t has to be endogenous to the interest variable y_t. Suited variables for Z_t are lagged values of y_t, as long as they are not included in Equation 4.1 and affect attrition likelihood. For attrition generated by unobservables, Z_t has to affect attrition and should not be related to y_t. In that case, good proxies for Z_t are any variables characterizing both interview process and interviewers. For example, Zabel (1998) suggests using the length of the interview, the number of contacts between the respondent and the interviewer, the number of adjustments that were made to certain variables, and the length of time the interviewer spent editing the response

forms. Other possibilities are variables indicating whether there was a change in the interviewer, or whether the interview was given to a proxy, or whether the interview was conducted by phone. Zabel (1998) expects these variables to be positively correlated with attrition.

The large majority of studies analyzing attrition focus on the selection on observables and leave behind the selection on unobservables (Alderman et al., 2001; Baigrie and Eyal, 2014; Outes-Leon and Dercon, 2009; Vaillant, 2013; Watson, 2003). The main challenge is finding appropriate proxies for the variables Z_t. They have to predict attrition but be independent of $\epsilon|X_t$ and not be included in the main interest equation. Maluccio (2004) is one of the rare authors to analyze attrition on unobservables. For Z_t variables, he uses information on the quality of the interview as an exclusionary restriction. He finds a non-random attrition related to the size and the resources of the household. Indicators of quality of the baseline interview significantly influence the probability to attrit. These indicators are used to correct for attrition bias due to sample selection on unobservables.

The second test comes from the four authors Becketti, Gould, Lillard and Welch (Becketti et al., 1988). The BGLW test can be seen as the inverse procedure of the FGM test (Outes-Leon and Dercon, 2009). The main equation of interest, Equation 4.1, is estimated by taking the baseline values of y and X (y_1 and X_1). The attrition dummy A_t is included as as explanatory variable, as well as an interaction between the attrition dummy and the other explanatory variables. The aim is to analyze if attritors differ in their initial characteristics. If interaction variables are significant, the estimates are likely to be subject to attrition bias (Maluccio et al., 2000). Then, the joint significance of A_t and the interaction variables is tested using a F-test. This allows an estimate as to whether the coefficients from the explanatory variables differ between stayers and attritors. Outes-Leon and Dercon (2009) see this test as a 'pooling test' because it compares the coefficients of the variables from the first wave between two samples : the full sample and the sample of stayers.

4.5.3 Correct for Attrition Bias

Estimates of y_t are considered as accurate if attrition tests did not provide evidence of attrition bias. Even if attrition is non-random, it does not necessarily bias the estimates (Outes-Leon and Dercon, 2009). But in the case of attrition bias, some solutions are suggested in the literature to correct estimates. Following Baigrie and Eyal (2014), I distinguish the solutions for selection on observables from the ones for selection on unobservables.

To correct bias coming from attrition on observables, Fitzgerald et al. (1998) suggest using weighted least squares (WLS) estimators. The approach is based on an inverse probability weighting (IPW) in two steps. First, the attrition equation is estimated using a probit estimator. Any variables

that affect attrition and are endogenous to y are included in Z_t.

$$S_t = \delta_0 + \delta_1 X_t + \delta_2 Z_t + \nu_t \tag{4.7}$$

$$S_t = \delta_0 + \delta_1 X_t + \mu_t \tag{4.8}$$

With $S_t = 1$ for stayers and $S_t = 0$ for attritors. Equation 4.8 is a restricted version of Equation 4.7. The variables included in Z_t must be observed for both stayers and attritors, and be correlated with attrition. For example, y_{t-1} can be a suitable z_t if it is related to attrition (Baigrie and Eyal, 2014). Baulch and Quisumbing (2011) also suggest that household demographic variables, community level variables, shocks or treatment variables can also be considered.

The ratio of predicted values from Equations 4.7 and 4.8 is then used to compute the weights. Finally, interest Equation 4.1 is estimated by WLS. The idea behind the procedure is to give more weight to stayers who have similar initial characteristics to attritors. This will give less importance to respondents whose characteristics make them more likely to remain in the panel. An advantage of this method is that it uses the full information available for the attritors (Falaris, 2003). Furthermore, the assumption required on Z_t is much weaker than in the selection model of Heckman : ϵ_t and ν_t should not be correlated.

However, Baulch and Quisumbing (2011) highlight some limitations of the method. First, attrition is assumed to come only from observable characteristics. In case this assumption does not hold because of selection on unobservables, Heckman's procedure is needed. Another is that this correction is model specific and should be completed for each interest variable.

For selection on unobservables, the correction procedure is the Heckman selection approach (Heckman, 1979). This method estimates a selection-corrected model in two steps. First, the attrition equation is estimated with all the exogenous variables x_t from the interest equation and the identified instruments z_t. These instruments must be related to attrition but not to the error term of the interest equation. Maluccio et al. (2000) use the variables measuring the quality of the first round interview. A selection correction factor or inverse of Mills' ratio (noted IMR below) estimated from the first step is then introduced into the second-stage estimation of the main interest equation.

$$y_t = \beta_0 + \beta_1 X_t + \beta_2 IMR + \epsilon_t \tag{4.9}$$

This approach is robust for attrition on both unobservables and observables (Maluccio et al., 2000).

One major difficulty is qualifying the exclusion restriction. The challenge is to find a suitable z_t which affects attrition but is preferably [4] not related to the interest variable (Fitzgerald et al., 1998;

4. In fact, it is not necessary for the selection equation to contain variables that are different from the explicative

Outes-Leon and Dercon, 2009). According to Fitzgerald et al. (1998); Zabel (1998) and Hill and Willis (2001), the best solution is to use the characteristics of both interviewers and the interview process as instruments. But these variables should not be related to respondent's characteristics. Maluccio et al. (2000) has a database including characteristics of the interview that are plausibly exogenous to a variety of outcomes. Therefore, he adopts the Heckman selection approach to correct for attrition bias. However, most of databases do not include such information and a large majority of articles on attrition focus on selection on observables.

4.6 RESULTS OF ATTRITION ANALYSIS

This section presents the results of attrition analysis. The first sub-section focuses on the characteristics of the attritors while the second presents the determinants of attrition in the database. The last sub-section discusses the limits of this analysis.

4.6.1 Characteristics of Attritors

A first step is to determine the characteristics of the total attritors, who represent 13.7% of the sample. The characteristics known for all respondents prior to the interview are the cohort and the region of SHS. We estimate the probability to be never interviewed in function of these pre-interview characteristics.

$$total\_attritor = \gamma_0 + \gamma_1 cohort\_FE + \gamma_2 region\_FE + \kappa_{clusterSHS} \tag{4.10}$$

Here, $total\_attritor$ equals 1 if the respondent is a total attritor and 0 if not. The other variables are defined in Chapter 3 (see Section C.2 in the Appendix for detailed definition of variables). Table 4.10 presents the results coming from three estimators : OLS, probit and logit. Belonging to a cohort other than C_{2011} is significantly and positively related to the probability to be a total attritor. However, the magnitude of the effect is the highest for the first cohort (C_{2005}) and decreases with the youth of the cohorts. The more the respondent comes from a young cohort, the lower is the probability to be a total attritor. This is explained by the fact that 80% of the total attritors come from the three first cohorts.

Moreover, being from a region other than Upper West has a significant and negative link with the probability to be a total attritor. This is consistent with what we found in the descriptive statistics part : retention rates are the highest in richest regions. One reason is that attendance is more likely

variables, although this is indeed desirable. The reason is that in this situation, identification is still possible because of the non-linearity of the Inverse Mills Ratio that avoids collinearity. However, this identification can be weak, especially when the distribution of X is not large.

higher in richer regions. Another is that regions further from Greater Accra have a lower probability to be well supervised as the leading team was based in the capital. This can have an impact on the attrition.

In a second step, the baseline characteristics of wave-attritors are analyzed. In particular, permanent stayers are compared to wave-attritors in Table 4.11. Column (1) presents the characteristics for the whole sub-sample. Column (2) reports means of wave-attritors, and Column (3) the ones of permanent stayers. Column (4) displays the t-test coefficients from differences in means. Standard errors are reported in Column (5). In order to control for cohort effects, the probability of attrition is also estimated in function of each variable separately and including cohort fixed effects. Columns (6) and (7) report coefficients and standard errors of OLS regressions.

Results suggest that there is no statistically significant gender difference. However, wave-attritors are younger than the permanent stayers (difference statistically significant). Some family characteristics are significantly different between the two groups. For example, wave-attritors have more siblings and their parents are less educated. Regarding education, wave-attritors have on average a higher aggregate score at Basic Education Certificate Examination (BECE) [5] than permanent stayers. The lower the BECE score, the better the grade. This therefore suggests that wave-attritors have poorer academics results than permanent stayers.

The survey does not collect information on the poverty level of the respondent's household. However, the level of education of the parents can be used as a proxy for household poverty. The results suggest that having a mother with no education does not change significantly the probability to be a permanent stayer or a wave-attritor. However, wave-attritors have a significantly higher probability to have a father with no education compared to permanent stayers. They also have a lower likelihood of having at least one parent with SHS-level education. This suggests that attritors are more likely to belong to poor household than stayers. For categorical variables such as SHS region, SHS subject and cohort, a Chi-square test is run. For these three variables, the null hypothesis is rejected, suggesting that all coefficients are not equal between stayers and attritors. However, it does not imply that there are differences between the two sub-samples for every category.

Table 4.12 presents the same tests as Table 4.11 but for baseline characteristics related to the labor market and asked in the 'Go Transition' survey. The sub-sample is restrained to cohorts C_{2005}, C_{2006} and C_{2007} having a 'Go Transition' baseline wave. [6] There is no statistically significant difference between wave-attritors and permanent stayers on migration characteristics such as being born in the current region and planning to stay there. Moreover, stayers and attritors do not significantly differ on labor market characteristics such as wage, worked hours, worked days, job search and expected

5. The BECE is taken at the end of the Junior High School and determines whether the student can go to SHS.
6. That is why the number of observations in this sub-sample is lower (2074 versus 4507 individuals).

TABLEAU 4.10 Total Attrition and Pre-Interview Characteristics

	(1) OLS	(2) Probit	(3) Logit
Cohorts			
Cohort 2005	0.361***	0.317***	0.330***
	(0.023)	(0.026)	(0.034)
Cohort 2006	0.265***	0.271***	0.285***
	(0.023)	(0.027)	(0.036)
Cohort 2007	0.070***	0.141***	0.144***
	(0.016)	(0.029)	(0.039)
Cohort 2008	0.004	0.033	0.048
	(0.010)	(0.028)	(0.038)
Cohort 2009	0.026***	0.063**	0.081**
	(0.009)	(0.027)	(0.036)
Cohort 2010	0.029***	0.067**	0.086**
	(0.010)	(0.027)	(0.036)
Regions			
Region Ashanti	-0.182***	-0.156***	-0.146***
	(0.037)	(0.026)	(0.029)
Region Brong Ahafo	-0.182***	-0.154***	-0.146***
	(0.038)	(0.026)	(0.030)
Region Central	-0.177***	-0.152***	-0.141***
	(0.035)	(0.024)	(0.028)
Region Eastern	-0.205***	-0.180***	-0.171***
	(0.034)	(0.023)	(0.026)
Region Greater Accra	-0.195***	-0.169***	-0.160***
	(0.034)	(0.024)	(0.027)
Region Northern	-0.236***	-0.213***	-0.209***
	(0.038)	(0.033)	(0.036)
Region Upper East	-0.199***	-0.168***	-0.164***
	(0.045)	(0.041)	(0.043)
Region Volta	-0.218***	-0.190***	-0.185***
	(0.036)	(0.027)	(0.031)
Region Western	-0.216***	-0.190***	-0.183***
	(0.037)	(0.029)	(0.033)
Constant	0.207***		
	(0.036)		
Observations	7616	7616	7616
R-squared	0.177		
Pseudo R-squared		0.206	0.195

Note : *** p<0.01, ** p<0.05, * p<0.1. Clustered at SHS level standard errors are in parentheses. Comparison categories are 'C_{2011}' for the cohort variable and 'Upper West' for the region variable. For the probit and logit estimations, marginal effects are presented instead of coefficients. C_{20XX} refers to the cohort of students that started SHS in 20XX. Anne Duplantier (2021)

TABLEAU 4.11 Differences in Characteristics Between Wave-Attritors and Permanent Stayers - Part 1

	(1) Full Sub-Sample	(2) Wave Attritors	(3) Permanent Stayers	(4) T-test Diff.	(5) T-test SE	(6) OLS Diff.	(7) OLS SE
Characteristics in 'In School' and 'Go Transition'							
Male (%)	58.54	58.41	58.57	-0.16	(2.05)	0.00	(0.01)
Age (average)	19.87	19.61	19.94	-0.33***	(0.09)	0.02***	(0.003)
	(0.02)	(0.06)	(0.03)				
Born in urban area (%)	48.33	41.79	49.85	-8.06***	(1.92)	-0.05***	(0.01)
At least one parent with SHS (%)	48.05	40.62	49.92	-9.3***	(1.99)	-0.06***	(0.01)
Father with no education (%)	3.69	5.4	3.28	2.12***	(0.74)	0.09***	(0.03)
Mother with no education (%)	9.11	9.73	8.96	0.77	(1.12)	0.01	0.02
Number of siblings (average)	3.97	4.21	3.92	0.29***	(0.08)	0.01***	(0.002)
	(0.03)	(0.08)	(0.04)				
Living with family/friend (%)	75.17	78.63	74.56	4.07*	(2.5)	0.01	(0.02)
Father died (%)	9.76	8.51	10.03	1.52	(1.19)	0.1***	(0.02)
Mother died (%)	2.93	3.17	2.88	0.28	(0.67)	0.18***	(0.02)
BECE aggregate score (average)	18.48	19.96	18.11	1.85***	(0.25)	0.01***	(0.00)
SHS region							
Ashanti (%)	16.50	13.77	17.18	χ^2 test p-value :		-0.08***	(0.02)
Brong Ahafo (%)	10.64	10.97	10.56	0.000		-0.04*	(0.03)
Central (%)	11.53	10.64	11.75			-0.07***	(0.02)
Eastern (%)	16.61	14.45	17.75			-0.08***	(0.02)
Greater Accra (%)	9.26	7.17	9.78			-0.09***	(0.03)
Northern (%)	5.38	6.61	5.07			0.02	(0.03)
Upper East (%)	2.33	4.14	1.88			0.12***	(0.04)
Upper West (%)	2.62	5.6	1.88			0.17***	(0.04)
Volta (%)	16.22	15.23	16.46			-0.07***	(0.02)
Western (%)	8.91	11.42	8.28			comparison categ.	
SHS subject							
Business (%)	25.13	23.29	25.58	χ^2 test p-value :		-0.02	(0.02)
General Arts (%)	24.97	25.53	24.83	0.017		-0.01	(0.02)
Science (%)	21.11	18.14	21.84			-0.03	(0.02)
Vocational (%)	0.36	0.56	0.3			0.10	(0.1)
Agriculture (%)	10.13	11.2	9.87			0.02	(0.02)
Visual Arts (%)	4.49	6.05	4.1			0.05	(0.03)
Technical (%)	3.80	4.59	3.6			0.05	(0.03)
Home Economics (%)	10.02	10.64	9.87			comparison categ.	
Cohort							
C_{2005} (%)	13.36	10.47	14.08	χ^2 test p-value :		-0.19***	(0.02)
C_{2006} (%)	15.11	9.24	16.57	0.000		-0.22***	(0.02)
C_{2007} (%)	17.55	9.91	19.45			-0.23***	(0.02)
C_{2008} (%)	17.11	14.81	17.68			-0.17***	(0.02)
C_{2009} (%)	9.85	11.02	9.56			-0.12***	(0.02)
C_{2010} (%)	10.76	16.48	9.34			-0.04*	(0.02)
C_{2011} (%)	16.26	28.06	13.33			comparison categ.	
In-school survey baseline (%)	53.98	70.38	49.9	20.47***	(1.83)	0.22***	(0.02)
2011 survey baseline (%)	83.73	71.94	86.67	-14.73***	(1.36)	-0.12***	(0.02)
Observations	4507	898	3609				

Note : *** p<0.01, ** p<0.05, * p<0.1. One observation per respondent. Only wave-attritors and non-attritors are considered here ($N = 4507$). C_{20XX} refers to the cohort of students that started SHS in 20XX. "SE" in columns (5) and (7) refers to standard errors. Anne Duplantier (2021)

probability to find a job in ten years. With regard to the ‘occupation of the parents’ and ‘primary activity of the respondent’ categorical variables, we cannot reject the null hypothesis that the two sub-samples are equal for every category. This suggests that there is no statistically significant difference between the two groups.

The third step sheds light on the last category of attritors : the respondents classified as temporary stayers, i.e. those who return to the survey after missing one or more waves. The baseline characteristics of temporary stayers are analyzed following the same analysis as for the wave-attritors. In Table 4.13, permanent stayers are compared to temporary stayers. Column (1) presents the characteristics for the whole sub-sample. Column (2) reports means of temporary stayers, and Column (3) those of permanent stayers. Column (4) displays the t-test coefficients from differences in means. Standard errors are reported in Column (5). In order to control for cohort effects, the probability of attrition is also estimated in function of each variable separately and including cohort fixed effects. Columns (6) and (7) report coefficients and standard errors of OLS regressions.

Unlike wave-attritors, there is a statistically significant gender difference between temporary and permanent stayers. Indeed, there are more male respondents in the group of temporary stayers than in the permanent stayers. As for wave-attritors, temporary stayers are slightly younger than permanent stayers (difference statistically significant). There is no significant difference (at 5% level) between the two groups in the proportion of people born in rural areas. Regarding the level of parents’ education, there is a statistically significant higher proportion of permanent stayers that have at least one parent with SHS-level education in the group of permanent stayers. However, there is no statistically significant difference in the proportion of respondents with at least one parent who has no education.

Regarding family characteristics, temporary stayers have on average a higher number of siblings (difference statistically significant). There is no statistically significant difference (at 5% level) between the two groups in the proportion of respondents living with family. Temporary stayers are less likely to have a dead mother or father (statistically significant), but this difference is very close to zero. Moreover, temporary stayers have on average higher BECE aggregate score, meaning that they have worse academic results (as explained previously).

For the three categorical variables - SHS region, SHS subject and cohort - the null hypothesis is rejected, suggesting that all coefficients are not equal between stayers and attritors. However, it does not mean that there are differences between the two sub-samples for every category. In general, OLS estimations (Columns 6 and 7) confirm results from the t-test.

Finally, Table 4.14 presents the baseline characteristics that are specific to the ‘Go Transition’ survey : occupation of parents, migration characteristics and labor market characteristics. There is no statistically significant difference between permanent and temporary stayers in the father’s oc-

TABLEAU 4.12 Differences in Characteristics Between Wave-Attritors and Permanent Stayers - Part 2

	(1) Full Sub-Sample	(2) Wave Attritors	(3) Permanent Stayers	(4) T-test Diff.	(5) T-test SE	(6) OLS Diff.	(7) OLS SE
Characteristics only in 'Go Transition'							
Father's usual occupation							
Farmer (%)	31.26	35.43	30.66	χ^2 test p-value :		0.15	(0.19)
Govnmt employee (%)	19.46	19.93	10.56	0.545		0.12	(0.19)
Teacher (%)	7.51	5.91	7.74			0.11	(0.19)
Self-empl (services) (%)	15.06	17.32	14.74			0.16	(0.19)
Self-empl (trading) (%)	10.07	9.45	10.16			0.13	(0.19)
Self-empl (manufact) (%)	3.16	2.76	3.22			0.12	(0.192)
Other wage empl (%)	13.33	12.99	13.38			0.14	(0.19)
Unemployed (%)	0.15	-	0.17			comparison categ.	
Mother's usual occupation							
Farmer (%)	20.41	22.61	20.09	χ^2 test p-value :		0.02	(0.05)
Govnmt employee (%)	4.6	3.07	4.83	0.867		-0.03	(0.06)
Teacher (%)	3.59	2.68	3.72			-0.02	(0.06)
Self-empl (services) (%)	4.7	4.6	4.72			0.01	(0.06)
Self-empl (trading) (%)	59.67	60.54	59.54			0.01	(0.05)
Self-empl (manufact) (%)	2.28	19.2	2.33			-0.01	(0.07)
Other wage empl (%)	2.67	2.68	2.66			0.01	(0.07)
Unemployed (%)	2.08	1.92	2.11			comparison categ.	
Born current region (%)	50.12	51.72	49.89	-1.83	(3.31)	0.01	(0.01)
Plan to stay current region (%)	88.4	86.59	88.66	-2.07	(2.12)	-0.02	(0.02)
Primary activity							
Education (%)	42.85	38.04	43.49	χ^2 test p-value :		-0.02	(0.05)
Work (%)	43.87	49.46	43.13	0.439		0.03	(0.05)
Unemployed (%)	10.34	9.78	10.42			0.03	(0.05)
Apprenticeship (%)	2.94	2.72	2.97			comparison categ.	
Weekly wage in GHS (average)	38.88	46.16	37.77	8.39	(6.8)	0.0001	(0.0002)
	(2.31)	(6.57)	(2.46)				
Weekly worked hours (average)	45.77	44.17	46	-1.83	(2.48)	-0.0001	(0.0006)
	(0.83)	(2.32)	(0.89)				
Weekly worked days (average)	5.19	5	5.22	-0.22	(0.16)	-0.01	(0.01)
	(0.05)	(0.17)	(0.05)				
Job search (%)	15.48	12.64	15.89	-3.25	(2.39)	-0.02	(0.02)
Expected prob. to find a job in 10 years							
After university (%)	83.76	83.16	83.85	-0.69	(1.25)	-0.0002	(0.0004)
No study after SHS (%)	55.86	54.86	56.01	-1.14	(1.82)	-0.0001	(0.0002)
Cohort							
C_{2005} (%)	29.03	35.34	28.1	χ^2 test p-value :		0.04**	(0.02)
C_{2006} (%)	32.84	31.20	33.08	0.045		0.01	(0.02)
C_{2007} (%)	38.14	33.46	38.83			comparison categ.	
Number of observations	2074	266	1808				

Note : *** p<0.01, ** p<0.05, * p<0.1. One observation per respondent. Only wave-attritors and stayers from the first three cohorts are considered here ($N = 2074$). C_{20XX} refers to the cohort of students that started SHS in 20XX. "SE" in columns (5) and (7) refers to standard errors. Anne Duplantier (2021)

TABLEAU 4.13 Differences in Characteristics Between Temporary and Permanent Stayers - Part 1

	(1) Full Sub-Sample	(2) Temp. Stayers	(3) Perm. Stayers	(4) T-test Diff.	(5) T-test SE	(6) OLS Diff.	(7) OLS SE
Characteristics in 'In School' and 'Go Transition'							
Male (%)	60.02	63.01	58.57	4.44***	(1.42)	0.05***	(0.01)
Age (average)	19.63	19.05	19.94	-0.89***	(0.09)	0.02***	(0.003)
	(0.03)	(0.06)	(0.04)				
Born in urban area (%)	49.02	47.39	49.85	-2.46*	(1.43)	-0.06***	(0.01)
At least one parent with SHS (%)	46.35	39.93	49.92	-9.99***	(1.49)	-0.1***	(0.01)
Father with no education (%)	3.5	3.9	3.28	0.62	(0.54)	0.01	(0.03)
Mother with no education (%)	9.04	9.18	8.96	0.22	(0.83)	-0.01	0.02
Number of siblings (average)	4.07	4.36	3.92	0.44***	(0.06)	0.01***	(0.002)
	(0.03)	(0.08)	(0.04)				
Living with family/friend (%)	75.45	82.42	74.56	7.86*	(2.85)	-0.02	(0.02)
Father died (%)	8.04	4.6	9.82	-0.05***	(0.01)	-0.001	(0.02)
Mother died (%)	2.17	0.81	2.88	-0.02***	(0.004)	-0.04***	(0.04)
BECE aggregate score (average)	18.47	19.18	18.11	1.07***	(0.19)	0.01***	(0.001)
SHS region							
Ashanti (%)	17.54	18.29	17.18	χ^2 test p-value :		0.05**	(0.02)
Brong Ahafo (%)	10.50	10.38	10.56	0.000		0.04	(0.03)
Central (%)	11.01	9.44	11.75			-0.01	(0.02)
Eastern (%)	14.98	10.38	17.75			-0.05**	(0.02)
Greater Accra (%)	9.5	8.91	9.78			0.02	(0.03)
Northern (%)	6.94	10.9	5.07			0.24***	(0.03)
Upper East (%)	3.61	7.27	1.88			0.35***	(0.03)
Upper West (%)	3.24	6.1	1.88			0.32***	(0.04)
Volta (%)	15.05	12.08	16.46			-0.02	(0.02)
Western (%)	7.64	6.27	8.28			comparison categ.	
SHS subject							
Business (%)	24.9	23.59	25.58	χ^2 test p-value :		0.03	(0.02)
General Arts (%)	24.17	22.89	24.83	0.000		0.03	(0.02)
Science (%)	21.31	20.28	21.84			0.03	(0.02)
Vocational (%)	0.35	0.43	0.3			0.05	(0.1)
Agriculture (%)	11.18	13.71	9.87			0.13***	(0.02)
Visual Arts (%)	4.62	5.6	4.1			0.11***	(0.03)
Technical (%)	4.2	5.55	3.6			0.17***	(0.03)
Home Economics (%)	9.21	7.95	9.87			comparison categ.	
Cohort							
C_{2005} (%)	10.16	3.3	14.08	χ^2 test p-value :		-0.29***	(0.02)
C_{2006} (%)	12.17	4.46	16.57	0.000		-0.27***	(0.02)
C_{2007} (%)	15.63	8.93	19.45			-0.2***	(0.02)
C_{2008} (%)	15.91	12.81	17.68			-0.11***	(0.02)
C_{2009} (%)	16.37	28.29	9.56			0.22***	(0.02)
C_{2010} (%)	15.47	26.2	9.34			0.21***	(0.02)
C_{2011} (%)	14.3	16.01	13.33			comparison categ.	
In-school survey baseline (%)	62.05	83.31	49.9	33.41***	(1.26)	0.29***	(0.02)
2011 survey baseline (%)	85.69	83.99	86.67	-2.68***	(0.01)	-0.27***	(0.02)
Observations	5670	2061	3609				

Note : *** p<0.01, ** p<0.05, * p<0.1. One observation per respondent. Only temporary and permanent stayers are considered here ($N = 5670$). "SE" in columns (5) and (7) refers to standard errors. Anne Duplantier (2021)

cupation of the respondent. However, the null hypothesis that all the categories of the mother's occupation are the same between the two groups is rejected. Regarding the migration characteristics, there is no statistically significant difference in the proportion of respondents born in the region they are currently living. However, more of the permanent stayers declare that they plan to stay in the region where they are currently living. This is consistent with what is expected from permanent stayers compared to temporary stayers. Indeed the latter could be more difficult to reach as they have a higher probability to move.

Almost all the characteristics regarding the labor market do not appear to be statistically significantly different between permanent and temporary stayers : primary activity, weekly wage, weekly worked hours, weekly worked days and the probability to search for a job. The two groups differ in their expected probability to find a job in ten years if they do not study after SHS. On average, this probability is statistically significantly higher for the permanent stayers than for the temporary stayers. However, the expected probability to find a job in ten years if they go to university after SHS does not differ between both groups. In general, OLS estimations (Columns 6 and 7) confirm results from the t-test.

In conclusion, this section allows us to describe the characteristics of the three types of attritors in comparisons of the permanent stayers. First, total attritors have a higher probability to belong to an older cohort. Moreover, the region in which they attended SHS is a statistically significant characteristic of total attritors. Second, wave attritors are younger, have less educated parents and more siblings than permanent stayers. They also have poorer academic results. There is, however, no statistically significant difference between both groups in the migration characteristics, labor market characteristics and occupation of their parents. Third, the differences between wave-attritors and permanent stayers are very similar to those between temporary and permanent stayers. However, the proportion of men is significantly higher within the temporary stayers group than the permanent stayers group. Moreover, temporary stayers have a lower probability of planning to stay in the region in which they are currently living.

4.6.2 Determinants of Attrition

This sub-section analyzes a potential attrition bias in education outcomes from Chapter 3 (application to university and attendance at university). Attrition coming from observables is analyzed first, followed by that from unobservables.

Attrition from Observables

To test for potential attrition bias due to observables, FGM and BGLW tests [7] are used. The FGM

7. These tests were previously presented in the methodology section.

TABLEAU 4.14 Differences in Characteristics Between Temporary and Permanent Stayers - Part 2

	(1) Full Sub-Sample	(2) Temporary Stayers	(3) Permanent Stayers	(4) T-test Diff.	(5) T-test SE	(6) OLS Diff.	(7) OLS SE
Characteristics only in 'Go Transition'							
Father's usual occupation							
Farmer (%)	30.73	31.51	30.66	χ^2 test p-value :		-0.18	(0.13)
Govnmt employee (%)	19.87	19.18	10.56	0.591		-0.18	(0.13)
Teacher (%)	7.51	4.79	7.74			-0.21	(0.13)
Self-empl (services) (%)	14.71	14.38	14.74			-0.18	(0.13)
Self-empl (trading) (%)	10.02	8.22	10.16			-0.2	(0.13)
Self-empl (manufact) (%)	3.29	4.11	3.22			-0.17	(0.13)
Other wage empl (%)	13.67	17.12	13.38			-0.16	(0.13)
Unemployed (%)	0.21	0.68	0.17			comparison categ.	
Mother's usual occupation							
Farmer (%)	20.21	21.62	20.09	χ^2 test p-value :		-0.02	(0.04)
Govnmt employee (%)	5.28	10.87	4.83	0.002		0.05	(0.05)
Teacher (%)	3.64	2.7	3.72			-0.04	(0.05)
Self-empl (services) (%)	4.67	4.05	4.72			-0.04	(0.05)
Self-empl (trading) (%)	58.72	48.65	59.54			-0.04	(0.04)
Self-empl (manufact) (%)	2.67	6.76	2.33			0.08	(0.05)
Other wage empl (%)	2.67	2.7	2.66			0.02	(0.05)
Unemployed (%)	2.15	2.7	2.11			comparison categ.	
Born current region (%)	49.64	46.67	49.89	-3.22	(4.25)	-0.01	(0.01)
Plan to stay current region (%)	88.25	83.33	88.66	-5.33**	(2.73)	-0.03*	(0.02)
Primary activity							
Education (%)	43.22	39.81	43.49	χ^2 test p-value :		-0.003	(0.04)
Work (%)	43.36	46.3	43.13	0.866		-0.02	(0.04)
Unemployed (%)	10.4	10.19	10.42			-0.04	(0.04)
Apprenticeship (%)	3.02	3.7	2.97			comparison categ.	
Weekly wage in GHS (average)	37.43	33.63	37.77	-4.13	(8.36)	-0.00004	(0.0002)
	(2.29)	(4.84)	(2.46)				
Weekly worked hours (average)	46.1	47.24	46	1.24	(3.09)	0.0002	(0.0005)
	(0.84)	(2.6)	(0.89)				
Weekly worked days (average)	5.22	5.46	5.22	0.24	(0.19)	0.01	(0.01)
	(0.05)	(0.17)	(0.05)				
Job search (%)	15.65	12.75	15.89	-3.14	(3.09)	-0.04**	(0.02)
Expected prob. to find a job in 10 years							
After university (%)	83.82	83.49	83.85	-0.36	(1.62)	-0.0001	(0.0004)
No study after SHS (%)	55.49	49.27	56.01	-6.74***	(2.34)	-0.001***	(0.0002)
Cohort							
C_{2005} (%)	26.77	19.77	28.1	χ^2 test p-value :		-0.09***	(0.02)
C_{2006} (%)	32.06	26.74	33.08	0.000		-0.07***	(0.02)
C_{2007} (%)	41.17	53.49	38.83			comparison categ.	
Number of observations	2152	344	1808				

Note : *** p<0.01, ** p<0.05, * p<0.1. One observation per respondent. Only temporary and permanent stayers from the first cohorts are considered here ($N = 2152$). "SE" in columns (5) and (7) refers to standard errors. Anne Duplantier (2021)

test estimates the probability of attrition in function of baseline values of X and y from the interest equation in Chapter 3. The purpose is to check whether the baseline values X_1 and y_1 predict the probability of attrition. The BGLW test estimates the baseline value of y in function of attrition and the usual X. The aim is to check whether y_1 is related to the probability of attrition. The sub-sample used in Chapter 3 contains individuals who have been interviewed simultaneously during an 'In School' wave (2012 or 2013 [8]) and a 'Go Transition' wave. The 'Inschool' questionnaire allows us to collect information about expectations regarding the labor market and post-secondary education, while the 'Go Transition' questionnaire gathers data about actual post-secondary education choices. The 'In School' wave is considered as the baseline and the 'Go Transition' wave as a follow-up survey.

The structure of the database brings some constraints. Attrition tests are based on baseline values of variables y and X. In Chapter 3, y is either application to university or attendance at university. Variables x are academic ability, expected returns to education and other individual characteristics such as age, gender, education of parents, number of siblings, being born in rural area. However, 'In School' waves do not gather information about baseline values of post-secondary education outcomes y_1, as respondents are still in SHS. There is no variation in y_1, nor in baseline value of ability. [9] As a consequence, I cannot run the BGLW test with the Chapter 3 sample.

This issue is addressed in two steps. First, the FGM test is run using baseline values of available explicative variables, i.e. expected return to education and control variables. Second, the BGLW test is run with a substitution sub-sample in order to overcome the limits of the Chapter 3 sub-sample. In that case, the 'Go Transition 2011' wave is considered as baseline survey. The substitution sub-sample is constituted from cohorts C_{2005}, C_{2006} and C_{2007} and is quite similar to Chapter 3 sub-sample. This offers the advantage of allowing us to test the presence of attrition bias as it gathers baseline values of y. The limits of the substitution sub-sample are discussed in the next section. The BGLW test estimates the baseline value of dependent variable y_1, i.e. application to and attendance at university, on the baseline values of explicative variables x_1, i.e. the probability to be an attritor and the interaction terms of explicative variables and attrition.

Table 4.15 presents the results of the FGM test using the Chapter 3 sub-sample. The estimated

8. The expectation questions are quite different for the 'In School 2011' wave, therefore I do not use this wave.
9. Ability is proxied by WASSCE score taken at the end of the SHS.

equation is :

$$\begin{aligned} wave\_attritor = \varphi_0 + \varphi_1 exp\_return\_uncondi + \varphi_2 exp\_maxmin + \\ \varphi_3 prcvd\_prob\_admitted\_uni + \varphi_4 prcvd\_prob\_afford\_uni \\ + \varphi_5 control\_var + \varphi_6 cohort\_FE + \varphi_7 region\_FE \\ + \kappa_{clusterSHS} \end{aligned} \tag{4.11}$$

where $wave\_attritor$ is a dummy variable that equals 1 if the respondent is a wave-attritor and 0 if not. The other variables are defined as in Chapter 3 (see Section C.2 in the Appendix for a detailed definition of variables). In Column (1), attrition likelihood is estimated with OLS. Results suggest that baseline values of explicative and control variables are not significantly related to the probability to be an attritor in the future waves. The only significant (at 10%) and positive variable is the probability to be born in a rural area. Column (2) is a robustness regression using a probit estimator. The results are very close. The only difference is that the 'rural born' variable is more significant i.e. at a level of 5% rather than 10%, with OLS estimator.

Table 4.16 presents findings of both FGM and BGLW tests run with the substitution sub-sample. Results of the FGM test from Equation 4.12 are presented in Columns (1) and (2).

$$\begin{aligned} wave\_attritor = \phi_0 + \phi_1 educ\_outcome + \phi_2 wasscescore + \phi_3 exp\_return\_uncondi \\ + \phi_4 exp\_maxmin + \phi_5 control\_var + \phi_6 cohort\_FE \\ + \phi_7 region\_FE + \rho_{clusterSHS} \end{aligned} \tag{4.12}$$

Here, $wave\_attritor$ is a dummy variable that equals 1 if the respondent is a wave-attritor and 0 if not. $educ\_outcome$ is either a dummy variable for application to university ($appl\_uni$) or a dummy variable for attendance at university (att_uni). The other variables are defined as in Chapter 3 (see Section C.2 in the Appendix for a detailed definition of variables).

There is a significant and negative relationship between application to university and probability to attrit. A respondent who applied to university in the past has less probability to become an attritor in the subsequent waves. There is also a significant and negative relationship between attrition and 'attendance at university' education outcome. A respondent who attended university in the past has less probability to become an attritor in the subsequent waves. This suggests that both the 'application to university' and 'attendance at university' outcomes of interest suffer from an attrition bias due to observables.

Results of BGLW test are presented in Columns (3) and (4) of Table 4.16. The aim is to analyze

TABLEAU 4.15 FGM Test of Wave Attrition - Chapter 3 sample

	(1) OLS	(2) Probit
Exp. uncondi. returns	-0.004	-0.004
	(0.009)	(0.008)
Exp. max-min	-0.006	-0.009
	(0.014)	(0.014)
Prob. enter uni	-0.000	-0.000
	(0.001)	(0.001)
Prob. finance uni	0.000	0.000
	(0.001)	(0.001)
Age	0.008	0.007
	(0.007)	(0.007)
Male	-0.006	-0.004
	(0.024)	(0.023)
Father primary educ	0.097	0.077*
	(0.062)	(0.046)
Father JHS educ	-0.026	-0.025
	(0.047)	(0.041)
Father SHS educ	0.004	0.007
	(0.045)	(0.040)
Father post-sec educ	-0.040	-0.041
	(0.050)	(0.045)
Mother primary educ	-0.015	-0.013
	(0.047)	(0.043)
Mother JHS educ	-0.008	-0.005
	(0.042)	(0.036)
Mother SHS educ	-0.001	-0.001
	(0.042)	(0.038)
Mother post-sed educ	0.001	0.004
	(0.053)	(0.054)
Born in rural area	0.045**	0.047**
	(0.019)	(0.019)
Number of siblings	0.002	0.001
	(0.005)	(0.005)
Constant	-0.054	
	(0.184)	
Region and cohort FE	Yes	Yes
Observations	1116	1116
R-squared	0.14	-
Pseudo R-squared	-	0.067

Note : *** p<0.01, ** p<0.05, * p<0.1. Clustered at SHS level standard errors are in parentheses. The dependent variable is the probability to be a wave-attritor. The marginal effects are presented for Probit estimation. Anne Duplantier (2021)

whether attritors differ in their initial characteristics from the stayers.

$$\begin{aligned} educ\_outcome = \theta_0 &+ \theta_1 wave\_attritor + \theta_2 wasscescore + \theta_3 wave\_attritor * wasscescore \\ &+ \theta_4 exp\_return\_uncondi + \theta_5 wave\_attritor * exp\_return\_uncondi \\ &+ \theta_6 exp\_maxmin + \theta_7 wave\_attritor * exp\_maxmin \\ &+ \theta_8 control\_var + \theta_9 wave\_attritor * control\_var \\ &+ \theta_1 0 cohort\_FE + \theta_1 1 region\_FE + \tau_{clusterSHS} \end{aligned} \tag{4.13}$$

Here, $wave\_attritor$ is a dummy variable that equals 1 if the respondent is a wave-attritor and 0 if not. $educ\_outcome$ is either a dummy variable for application to university ($appl\_uni$) or a dummy variable for attendance at university (att_uni). The other variables are defined as in Chapter 3 (see Section C.2 in the Appendix for a detailed definition of variables). Because of its length, Table 4.16 is truncated with only the main variables. The full table is available in Section C.3 in the Appendix in Table C.3.

The attrition dummy is not significantly related to the baseline value of application to university. However, there is a significant and negative relationship between the attrition dummy and the baseline value of attendance at university. This suggests a bias for attendance at university. The other variables of interaction between the attrition dummy and the other characteristics are not significantly related to both education outcomes. [10] Finally, I test the joint significance of the attrition variable and the interaction variables with F-tests derived from Wald tests. For both post-secondary education outcomes, the coefficient of baseline values are different for the full sample and the sample without attrition.

In conclusion, the FGM test with the original sample suggests no attrition bias. However, the FGM test using the substitution sample identifies a potential attrition bias coming from observables. The BGLW test suggests an attrition bias only for attendance. However, one cannot be sure that there is no attrition bias, as application to and attendance at university are negatively and significantly related to the probability to attrit in the substitution sub-sample. These potential biases will be addressed and corrected in Section 7.

As a robustness check, the same tests are run using a probit estimator. Results are presented in Table 4.17 and are globally similar to the main estimations. The only difference comes from the BGLW test for the attendance at university in Column (4). The relationship between the dummy of attrition and the baseline value of attendance is statistically significant at the level of 10% only. Because of

10. The few exceptions concern the interaction between attrition and the gender, as well as the education of the mother. These interaction variables have a significant and negative relationship with the baseline value of application to university.

TABLEAU 4.16 Tests of Wave Attrition - Substitution sample - OLS

	FGM test		BGLW test (Baseline value of)	
Dependent var.	Wave-attrition		Application	Attendance
	(1)	(2)	(3)	(4)
Applied to university	-0.037**			
	(0.014)			
Attended university		-0.045***		
		(0.012)		
Wave-attritors			-0.115	-0.555**
			(0.404)	(0.253)
Actual WASSCE score	0.001	0.001	-0.017***	-0.022***
	(0.001)	(0.001)	(0.002)	(0.002)
Attrit*WASSCE_score			0.000	0.005
			(0.004)	(0.003)
Exp. uncondi returns	0.009		0.005	
	(0.009)		(0.007)	
Attrit*return			0.000	
			(0.000)	
Exp. max-min	0.006		0.010	
	(0.012)		(0.016)	
Attrit*max_min			-0.032	
			(0.040)	
Constant	0.090	0.074	0.889***	1.166***
	(0.122)	(0.098)	(0.177)	(0.131)
Region and cohort FE	Yes	Yes	Yes	Yes
Control variables	Yes	Yes	Yes	Yes
Attrit*control_var	No	No	Yes	Yes
Observations	1737	2439	1737	2439
R-squared	0.027	0.023	0.215	0.213
Wald test			F(16,135) = 3.37	F(14,135) = 2.46
			P-value = 0.0001	P-value = 0.004

Note : *** p<0.01, ** p<0.05, * p<0.1. Clustered at SHS level standard errors are in parentheses. Because of its length, Table 4.16 is truncated with only the main variables. The full table is available in Section C.3 of the Appendix, in Table C.3. Anne Duplantier (2021)

its length, Table 4.17 is truncated with only the main variables. The full table is available in Section C.3 of the Appendix, in Table C.4.

TABLEAU 4.17 Tests of Wave Attrition - Substitution sample - Probit

	FGM test		BGLW test (Baseline value of)	
Dependent var.	Wave-attrition		Application	Attendance
	(1)	(2)	(3)	(4)
Applied to university	-0.039***			
	(0.015)			
Attended university		-0.050***		
		(0.014)		
Wave-attritors			0.259	-0.536*
			(0.602)	(0.290)
Actual WASSCE score	0.001	0.000	-0.016***	-0.022***
	(0.001)	(0.001)	(0.002)	(0.002)
Attrit*WASSCE_score			-0.013	-0.010
			(0.008)	(0.007)
Exp. uncondi returns	-0.000		-0.000	
	(0.000)		(0.000)	
Attrit*returns			0.000*	
			(0.000)	
Exp. max-min	0.004		0.011	
	(0.011)		(0.015)	
Attrit*max_min			-0.055	
			(0.058)	
Region and cohort FE	Yes	Yes	Yes	Yes
Control variables	Yes	Yes	Yes	Yes
Attrit*control_var	No	No	Yes	Yes
Observations	1737	2439	1728	2439
Pseudo R-squared	0.046	0.041	0.188	0.199
Wald test			Chi2(15) = 45.97	Chi2(14) = 37.85
			P-value = 0.0001	P-value = 0.0005

Note : *** p<0.01, ** p<0.05, * p<0.1. Clustered at SHS level standard errors are in parentheses. Marginal effects are displayed. Because of its length, Table 4.17 is truncated with only the main variables. The full table is available in Section C.3 of the Appendix, in Table C.4. Anne Duplantier (2021)

Attrition from Unobservables

For analyzing attrition coming from unobservables, one needs variables z that characterize the interview. Precisely, our database captures the interview's length and the number of waves in which the enumerator participated. These characteristics are potentially related to attrition but are less likely to be related to post-secondary education outcomes. Therefore, they seem to be good proxies for capturing attrition coming from unobservables. However, a potential threat is that the interview's length may be related to the respondent's ability, which is linked to post-secondary education outcomes. Moreover, questionnaire content (and thus length) varies according to the respondent's characteristics. For example, the section on application will be skipped for a respondent who has never applied to any post-secondary institution. These assumptions will be checked in Tables 4.19 and 4.20. Descriptive statistics are presented in Tables 4.18 and 4.19 before the attrition test is run.

As explained in Chapter 1, we tried to hire the same surveyors in order to ensure the quality of the data collection. Some surveyors were present during all the waves of the survey while other participated in only one. Table 4.18 reports the proportion of respondents having been interviewed by a surveyor that worked in one or more waves. Column (1) shows that 43.60% of the observations have been collected by enumerators who have been involved in the project for only one wave. It means that for 57% of the observations, the enumerators have participated in at least two waves of our survey. The remaining columns present the statistics disaggregated by survey wave. This gives an indication of the capacity of enumerators, as the good enumerators were re-hired for the subsequent waves. We expect that having a high proportion of experienced enumerators has a positive impact on the quality of the data and decreases the attrition rate.

TABLEAU 4.18 Number of Waves in which Enumerator Participated

	All	IN11	GOT11	IN12	GOT12	IN13	GOT14	GOT18
1 wave	43.60	44.06	37.25	44.21	41.36	60.81	28.66	58.71
2 waves	12.85	9.16	15.23	9.74	10.36	3.28	20.44	12.67
3 waves	17.16	31.25	16.15	17.43	22.59	18.93	11.10	12.26
4 waves	8.25	1.50	12.13	8.18	18.15	2.91	12.01	0.00
5 waves	15.56	14.03	18.44	20.40	7.54	9.90	23.07	11.16
7 waves	2.58	0.00	0.18	0.03	0.00	4.17	4.71	5.20
Observations	30464	3264	3264	3264	4352	3264	7616	5440

Note : Percentage of observations having been interviewed by enumerator having participated from one to all waves. INXX refers to the wave 'In School' 20XX and GOTXX refers to the wave 'Go Transition' 20XX. Anne Duplantier (2021)

Table 4.19 presents average durations of interview in minutes. The length of interview is important because it can impact attrition. Indeed, if the interview is too long, the respondent may be discouraged to participate in the subsequent waves of interview. On average, the interview duration is 34 minutes (Panel A). Panel B disaggregates interview's length by wave of the survey. 'Go Transition'

surveys of 2011 and 2012 are the longest with 76 and 67 minutes, while 'In School' waves are the shortest : between 18 and 21 minutes.

Panel C presents the length of interview by education outcomes. This allows us to check the assumptions discussed earlier regarding the limits of this choice of variable. Line (a) shows statistics for respondents who did not apply to university, Line (b) gives the same for individuals who applied to university and Line (c) presents the difference in coefficient between the two groups of respondents and the significant level of the t-test associated. There is a statistically significant difference in interview's length between respondents who applied to university and those who did not. The same pattern is observed for respondents who attend university (Lines (d) to (f)). This confirms our assumption that interview's length might be correlated with education outcomes.

TABLEAU 4.19 Duration of the Interview

	Obs	Mean	Std. Dev	Min	Max
		Panel A : Whole survey			
All Sample	17990	34.22	25.98	4.2	119.98
		Panel B : By survey			
In School 2011	2710	21.86	9.53	5.63	68.38
Go Transition 2011	2098	76.12	16.92	28.36	119.98
In School 2012	986	18.00	6.53	5.81	70.66
Go Transition 2012	2862	67.16	18.10	22.93	119.15
Go Transition 2014	5166	21.45	10.44	4.31	117.32
Go Transition 2018	4168	28.18	9.50	4.2	117.6
		Panel C : By education outcomes			
(a) No application to uni	11483	33.11	25.22	4.2	119.95
(b) Application to uni	6507	36.17	27.16	4.26	119.98
(c) Difference (a) - (b)		-3.06***	0.40		
(d) No attendance at uni	13592	33.49	25.45	4.2	119.98
(e) Attendance at uni	4398	36.45	27.45	4.26	119.3
(f) Difference (d) - (e)		-2.96***	0.45		

Note : *** p<0.01, ** p<0.05, * p<0.1. Time is in minutes. During the wave 'In School 2013', the time variable has not been saved, so information is missing. Anne Duplantier (2021)

Checking this assumption another way, Table 4.20 shows results of OLS regressions of education outcomes (application to university and attendance at university) using interview length as explicative variable. In order to control for cohort effects, estimations include cohort fixed effects and cluster at the SHS level. There is a positive (though almost null) and statistically significant (at 5% level) relationship between the probability to apply to university and the time of interview. However, the relationship between the probability to attend university and the length of interview is not statistically significant. In light of these results, it seems that the correlation between education

outcomes and interview's length is weak. Therefore, interview length can be considered as a good z along with the number of interviews done by the surveyor.

TABLEAU 4.20 Correlation Between Education Outcomes and Interview's Length

	(1) Application to university	(2) Attendance at university
Time of interview	0.0004**	0.00004
	(0.0001)	(0.0001)
Cohort		
C_{2005}	0.12***	0.13***
	(0.02)	(0.01)
C_{2006}	0.11***	0.10***
	(0.01)	(0.01)
C_{2007}	0.17***	0.14***
	(0.01)	(0.01)
C_{2008}	0.18***	0.14***
	(0.01)	(0.01)
C_{2009}	0.12***	0.08***
	(0.01)	(0.01)
C_{2010}	0.07***	0.04***
	(0.01)	(0.01)
Constant	0.22***	0.14***
	(0.01)	(0.01)

Note : *** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at SHS level and are in parentheses. Anne Duplantier (2021)

Table 4.21 summarizes results of the FGM test estimated with two z_t variables characterizing the process of interview : length of interview and enumerator's number of participation. The attrition test is run with OLS and with the substitution sample presented earlier.

$$\begin{aligned} wave\_attritor = \sigma_0 &+ \sigma_1 time\_interview + \sigma_2 participation\_enum + \sigma_3 educ\_outcome \\ &+ \sigma_4 wasscescore + \sigma_5 exp\_return\_uncondi + \sigma_6 exp\_maxmin \\ &+ \sigma_7 control\_var + \sigma_8 cohort\_FE + \sigma_9 region\_FE + \zeta_{clusterSHS} \end{aligned} \tag{4.14}$$

Here, $wave\_attritor$ is a dummy variable that equals 1 if the respondent is a wave-attritor and 0 if not. $time\_interview$ equals the number of minutes the interview lasted and $participation\_enum$ equals the number of times that the enumerator interviewing the respondent participated in a wave of our survey. $educ\_outcome$ is either a dummy variable for application to university ($appl\_uni$) or a dummy variable for attendance at university (att_uni). The other variables are defined as in

Chapter 3 (see Section C.2 of the Appendix for a detailed definition of variables).

Column (1) reports results for application to university while the second column focuses on attendance at university. In both cases, variables characterizing the interview do not have a significant relationship with likelihood of attrition. However, as for attrition coming from observables, the likelihood to apply to university is significantly and positively associated with attrition. This suggests that there might be an attrition bias from unobservable characteristics while estimating the education outcomes.

TABLEAU 4.21 FGM Test of Wave Attrition with Unobservables - OLS

	(1)	(2)
	Substitution sample	
Time of interview	-0.001	-0.001
	(0.001)	(0.001)
Enumerator's participation	-0.003	-0.003
	(0.004)	(0.004)
Applied to university	-0.045***	
	(0.015)	
Attended university		-0.043***
		(0.012)
Actual wassce score	0.000	0.000
	(0.001)	(0.001)
Exp. uncondi returns	-0.000	
	(0.000)	
Exp. max-min	0.006	
	(0.011)	
Age	0.002	0.004
	(0.005)	(0.004)
Male	-0.002	-0.005
	(0.016)	(0.012)
Father primary educ	0.060	0.071*
	(0.046)	(0.037)
Father JHS educ	-0.007	-0.010
	(0.027)	(0.020)
Father SHS educ	0.031	0.019
	(0.030)	(0.024)
Father post-sec educ	0.006	0.002
	(0.030)	(0.023)
Mother primary educ	-0.020	-0.008
	(0.030)	(0.024)
Mother JHS educ	-0.009	0.004
	(0.021)	(0.018)
Mother SHS educ	-0.027	0.008
	(0.027)	(0.022)
Mother post-sec educ	-0.008	0.020
	(0.032)	(0.027)
Born in rural area	0.001	0.010
	(0.015)	(0.013)
Number of siblings	-0.003	-0.002
	(0.003)	(0.003)
Constant	0.099	0.072
	(0.136)	(0.107)
Region and cohort FE	Yes	Yes
Observations	1681	2361
R-squared	0.031	0.026

Note : *** p<0.01, ** p<0.05, * p<0.1. Clustered at SHS level standard errors are in parentheses. The dependent variable is the probability to be a wave attritor. The estimator is OLS. Anne Duplantier (2021)

4.6.3 Limits

Using a substitution sub-sample in order to analyze the consequences of attrition has some limits. The substitution sample from 'Go Transition 2011' baseline has the advantage of including simultaneously post-secondary education outcomes and expectations about the labor market. However, some information is missing. In particular, for testing attrition on 'application to university' education outcome, there is no information about expected academic ability and expected financial ability. Moreover, for testing attrition on 'attendance at university' education outcome, all labor market expectations variables and expected academic and financial ability variables are missing. Therefore, it is not possible to run attrition tests using all variables included in the interest equation from Chapter 3.

In addition, there are some differences between the two sub-samples. Substitution respondents come from cohorts C_{2005}, C_{2006} and C_{2007} whereas Chapter 3 respondents come from cohorts C_{2009}, C_{2010} and C_{2011}. One key difference between the two sub-samples is the age at which respondents were surveyed. Respondents from the substitution sample were slightly older than respondents from the original sample. The question is to know whether the link between attrition and education outcomes is different for older respondents. If we assume that older students are more mobile, the risk of losing respondents from the substitution sample is higher than in the original sample and it can impact the attrition results. Therefore, using the substitution sample might lead to an over-estimation of attrition bias if mobility is correlated with outcomes.

Another difference is that the baseline survey of the substitution sub-sample was collected out of school, whereas the baseline wave of the Chapter 3 sample was gathered at school. Moreover, we know from earlier sections that the first cohorts have more attritors than others. As a consequence, attrition is higher in the substitution sub-sample than in the Chapter 3 sub-sample. This can have an impact on attrition results. In particular, attrition bias can be higher for the substitution sub-sample than for the interest sub-sample. Therefore the attrition bias could be overestimated.

Despite these limits, the substitution sub-sample has some interesting advantages and is quite similar to the sub-sample of interest. The two sub-samples come from the same database and have the same regional representation : respondents come from the same region and the same SHS. The only difference is that they do not come from the same cohorts. There is no reason a priori why older respondents have a different relationship between education outcomes and attrition than the younger ones. It will therefore be useful to study the presence of attrition bias for the substitution sub-sample, as it could bring us information about a potential attrition bias in the database.

4.7 CORRECTION FOR ATTRITION BIAS

In this section, I correct the estimations of education outcomes from Chapter 3 (the probability to apply to university and the probability to attend university) for attrition bias. Two methods are employed : WLS estimation and the Heckman procedure. The WLS estimator is proposed by Fitzgerald et al. (1998) to correct attrition bias from observables.

In a first step, the IPW is computed for each interest variable. The weight is the ratio of predicted values from the estimations of attrition equation (Equation 4.15) and restricted attrition equation (Equation 4.16) using a probit estimator.

$$\begin{aligned} A = \delta_0 &+ \delta_1 educ\_outcome + \delta_2 wasscescore + \delta_3 exp\_return\_uncondi + \delta_4 exp\_maxmin \\ &+ \delta_5 control\_var + \delta_6 time\_interview + \delta_7 participation\_enum \\ &+ \delta_8 cohort\_FE + \delta_9 region\_FE + \nu_{clusterSHS} \end{aligned} \tag{4.15}$$

$$\begin{aligned} A = \alpha_0 &+ \alpha_1 educ\_outcome + \alpha_2 wasscescore + \alpha_3 exp\_return\_uncondi + \alpha_4 exp\_maxmin \\ &+ \alpha_5 control\_var + \alpha_6 cohort\_FE + \alpha_7 region\_FE + \mu_{clusterSHS} \end{aligned} \tag{4.16}$$

Here, $A = 1$ if the respondent is an attritor and $A = 0$ if the respondent is a stayer. $educ\_outcome$ is either a dummy variable for application to university ($appl\_uni$) or a dummy variable for attendance at university (att_uni). $time\_interview$ equals the number of minutes the interview lasted and $participation\_enum$ equals the number of times that the enumerator interviewing the respondent participated in a wave of our survey. All the other variables are defined as in Chapter 3 and details are reported in Section C.2 of the Appendix.

Then, interest equation from Chapter 3 is re-estimated using the WLS estimator (see Section C.2 of the Appendix for details on variables).

$$\begin{aligned} educ\_outcome = \beta_1 exp\_return\_uncondi + \beta_2 exp\_maxmin + \beta_3 wasscescore \\ + \beta_4 control\_var + \beta_5 cohort\_FE + \beta_6 region\_FE + \epsilon_{clusterSHS} \end{aligned} \tag{4.17}$$

Results from the WLS estimation are presented in Tables 4.22 and 4.23. Results in Table 4.22 correspond to results in Tables 10 and 11 of Chapter 3, while results in Table 4.23 are compared to results in Tables 12 and 13 of Chapter 3.

There are no major differences between estimations without and with correction. Almost all the

coefficients have the same sign, magnitude [11] and significativity level. One slight difference lies in Columns (1) and (3) of Table 4.22. The coefficients of expected unconditional returns to education are no longer significant. Another difference relates to the control variable in Column (1) of Table 4.22. After correction, respondents whose parent (father or mother) has a post-secondary level of education have a statistically significant higher probability to apply to university than respondents whose parent has no education. The impact is larger for mother than father.

TABLEAU 4.22 Corrected Estimation - WLS

	Application				Attendance			
	Corrected	Chap3	Corrected	Chap3	Corrected	Chap3	Corrected	Chap3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Exp. uncondi. returns	2.03e-06	0.013**	-5.76e-06	-0.007	1.71e-06	0.009*	3.19e-06	-0.007
	(4.89e-06)	(0.005)	(6.45e-06)	(0.010)	(5.70e-06)	(0.005)	(7.08e-06)	(0.009)
Exp. max-min	0.017	0.004	0.023	0.019	0.032**	0.015*	0.015	0.028**
	(0.015)	(0.010)	(0.017)	(0.015)	(0.015)	(0.009)	(0.018)	(0.014)
Actual wassce score			0.009***	0.019			0.008***	0.007***
			(0.001)	(0.015)			(0.001)	(0.001)
Prcvd prob. admitted uni			0.001	0.001			0.001	0.001
			(0.001)	(0.001)			(0.001)	(0.001)
Prcvd prob. afford uni			0.001**	0.002***			0.001	0.001*
			(0.001)	(0.001)			(0.001)	(0.001)
Age	-0.027***	-0.033***	-0.015*	-0.017**	-0.026***	-0.024***	-0.018**	-0.016**
	(0.007)	(0.005)	(0.008)	(0.008)	(0.007)	(0.004)	(0.008)	(0.006)
Male	0.159***	0.140***	0.112***	0.112***	0.117***	0.089***	0.079***	0.068**
	(0.027)	(0.024)	(0.033)	(0.033)	(0.024)	(0.020)	(0.028)	(0.027)
Father primary educ	-0.056	-0.039	-0.018	-0.067	-0.067	-0.032	0.001	-0.023
	(0.059)	(0.044)	(0.077)	(0.069)	(0.053)	(0.033)	(0.071)	(0.055)
Father JHS educ	-0.050	-0.029	-0.112**	-0.113**	-0.064*	-0.024	-0.080	-0.072*
	(0.040)	(0.029)	(0.053)	(0.049)	(0.037)	(0.022)	(0.050)	(0.041)
Father SHS educ	-0.026	-0.018	-0.069	-0.075	-0.022	-0.010	-0.039	-0.030
	(0.039)	(0.031)	(0.051)	(0.048)	(0.039)	(0.026)	(0.050)	(0.044)
Father post-sec educ	0.088**	0.073**	-0.026	-0.034	0.050	0.041	-0.000	-0.015
	(0.038)	(0.031)	(0.053)	(0.048)	(0.038)	(0.027)	(0.055)	(0.045)
Mother primary educ	0.071*	0.060**	0.008	0.027	0.053	0.030	-0.003	0.009
	(0.042)	(0.030)	(0.049)	(0.048)	(0.041)	(0.026)	(0.051)	(0.045)
Mother JHS educ	0.045	0.022	0.013	0.024	0.053*	0.019	0.017	0.010
	(0.033)	(0.023)	(0.046)	(0.042)	(0.031)	(0.019)	(0.040)	(0.032)
Mother SHS educ	0.032	0.033	-0.001	-0.009	0.057	0.054*	0.029	0.030
	(0.040)	(0.031)	(0.051)	(0.047)	(0.037)	(0.028)	(0.047)	(0.042)
Mother post-sec educ	0.140***	0.103**	0.055	0.035	0.069	0.049	-0.011	-0.024
	(0.050)	(0.042)	(0.063)	(0.061)	(0.044)	(0.032)	(0.053)	(0.045)
Number of siblings	-0.011**	-0.008**	-0.012**	-0.009**	-0.008*	-0.005*	-0.008*	-0.004
	(0.005)	(0.004)	(0.005)	(0.004)	(0.004)	(0.003)	(0.005)	(0.004)
Born in rural area	-0.033	-0.016	-0.036	-0.015	-0.008	-0.007	-0.023	-0.018
	(0.023)	(0.018)	(0.029)	(0.027)	(0.022)	(0.016)	(0.028)	(0.025)
Constant	0.667***	0.747***	-0.129	-0.196	0.446**	0.400***	-0.076	-0.239
	(0.188)	(0.139)	(0.223)	(0.236)	(0.184)	(0.110)	(0.223)	(0.214)
Region fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2212	3314	1267	1323	2212	3314	1267	1323
R-squared	0.121	0.091	0.212	0.198	0.106	0.076	0.196	0.171

Note : *** p<0.01, ** p<0.05, * p<0.1. Clustered at SHS level standard errors are in parentheses. Anne Duplantier (2021)

The second method used to correct for attrition bias is the Heckman procedure (Heckman, 1979).

11. The coefficient is often slightly larger after correction than without correction.

TABLEAU 4.23 Corrected Estimation by gender - WLS

	Application				Attendance			
	Men		Women		Men		Women	
	Corrected	Chap3	Corrected	Chap3	Corrected	Chap3	Corrected	Chap3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Exp. uncondi. returns	-6.70e-06	-0.016	-6.61e-06	0.001	1.58e-06	-0.007	2.98e-06	-0.011
	(7.37e-06)	(0.012)	(1.12e-05)	(0.014)	(8.47e-06)	(0.012)	(1.14e-05)	(0.014)
Exp. max-min	0.019	0.013	0.044*	0.042*	0.016	0.029	0.027	0.038*
	(0.024)	(0.021)	(0.026)	(0.021)	(0.024)	(0.019)	(0.027)	(0.021)
Actual wassce score	0.011***	0.011***	0.007***	0.007***	0.010***	0.009***	0.006***	0.005***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Prcvd prob. admitted uni	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Prcvd prob. afford uni	0.001	0.001*	0.002**	0.002***	0.001	0.001	0.001	0.001*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Age	-0.010	-0.015*	-0.023	-0.014	-0.010	-0.013*	-0.034**	-0.019
	(0.009)	(0.009)	(0.016)	(0.013)	(0.009)	(0.008)	(0.014)	(0.011)
Father primary educ	0.034	0.004	-0.145	-0.210*	0.037	0.006	-0.064	-0.083
	(0.102)	(0.092)	(0.135)	(0.125)	(0.086)	(0.069)	(0.133)	(0.095)
Father JHS educ	-0.163**	-0.125*	-0.060	-0.102	-0.111	-0.111**	-0.057	-0.025
	(0.072)	(0.064)	(0.080)	(0.087)	(0.067)	(0.053)	(0.080)	(0.070)
Father SHS educ	-0.072	-0.058	-0.095	-0.108	-0.024	-0.032	-0.094	-0.042
	(0.068)	(0.060)	(0.077)	(0.085)	(0.061)	(0.050)	(0.090)	(0.081)
Father post-sec educ	0.030	0.043	-0.108	-0.127	0.042	-0.012	-0.071	-0.020
	(0.077)	(0.068)	(0.076)	(0.083)	(0.081)	(0.062)	(0.082)	(0.074)
Mother primary educ	-0.005	0.012	0.051	0.060	0.000	0.005	0.013	0.019
	(0.067)	(0.068)	(0.074)	(0.073)	(0.069)	(0.058)	(0.066)	(0.063)
Mother JHS educ	0.038	0.036	0.020	0.032	0.039	0.029	0.013	-0.001
	(0.057)	(0.052)	(0.064)	(0.066)	(0.051)	(0.041)	(0.052)	(0.048)
Mother SHS educ	-0.006	-0.005	0.051	0.013	0.045	0.054	0.040	0.019
	(0.071)	(0.068)	(0.068)	(0.062)	(0.065)	(0.059)	(0.062)	(0.058)
Mother post-sec educ	0.017	0.033	0.106	0.043	-0.033	0.000	0.018	-0.049
	(0.086)	(0.090)	(0.080)	(0.080)	(0.081)	(0.066)	(0.082)	(0.069)
Number of siblings	-0.014*	-0.009	-0.006	-0.005	-0.010	-0.005	-0.006	-0.003
	(0.008)	(0.007)	(0.007)	(0.006)	(0.008)	(0.007)	(0.007)	(0.006)
Born in rural area	-0.016	-0.001	-0.056	-0.024	-0.003	-0.004	-0.040	-0.030
	(0.033)	(0.032)	(0.044)	(0.038)	(0.033)	(0.030)	(0.043)	(0.038)
Constant	-0.116	-0.072	-0.096	-0.364	-0.181	-0.257	0.178	-0.188
	(0.284)	(0.286)	(0.361)	(0.336)	(0.278)	(0.278)	(0.332)	(0.305)
Region fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	755	803	512	520	755	803	512	520
R-squared	0.233	0.208	0.195	0.191	0.218	0.186	0.174	0.162

Note : *** p<0.01, ** p<0.05, * p<0.1. Clustered at SHS level standard errors are in parentheses. Anne Duplantier (2021)

In a first step, the attrition equation is estimated using a probit estimator and with two instruments : length of interview and enumerator experience.

$$\begin{aligned} A = \delta_0 &+ \delta_1 educ\_outcome + \delta_2 wasscescore + \delta_3 exp\_return\_uncondi + \delta_4 exp\_maxmin \\ &+ \delta_5 control\_var + \delta_6 time\_interview + \delta_7 participation\_enum \\ &+ \delta_8 cohort\_FE + \delta_9 region\_FE + \vartheta_{clusterSHS} \end{aligned} \tag{4.18}$$

All the variables are the same as in the WLS method (see Section C.2 of the Appendix for details on variables). The inverse of Mill's ratio is predicted from Equation 4.13. This selection correction factor is then introduced in the main interest equation from Chapter 3. The second and last step of the Heckman procedure is the estimation of Equation 4.14 with OLS (see Section C.2 of the Appendix for details on variables).

$$\begin{aligned} educ\_outcome = \lambda_1 exp\_return\_uncondi + \lambda_2 exp\_maxmin + \lambda_3 wasscescore + \lambda_4 control\_var \\ + \lambda_5 inverse\_mills\_ratio + \lambda_6 cohort\_FE + \lambda_7 region\_FE + \eta_{clusterSHS} \end{aligned} \tag{4.19}$$

Results from the Heckman correction are presented in Tables 4.24 and 4.25. These are very similar to results from weighted estimation and results from Chapter 3.

Finally, corrected estimations are very close to results presented in Chapter 3, regardless of the correction method used. We can conclude that results in Chapter 3 are robust and not strongly altered by attrition bias.

TABLEAU 4.24 Corrected Estimation - Heckman

	Application				Attendance			
	Corrected	Chap3	Corrected	Chap3	Corrected	Chap3	Corrected	Chap3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Exp. uncondi. returns	2.30e-06	0.013**	-6.22e-06	-0.007	2.34e-06	0.009*	-9.49e-07	-0.007
	(4.15e-06)	(0.005)	(6.17e-06)	(0.010)	(3.58e-06)	(0.005)	(5.56e-06)	(0.009)
Exp. max-min	0.004	0.004	0.023	0.019	0.015	0.015*	0.022	0.028**
	(0.013)	(0.010)	(0.019)	(0.015)	(0.011)	(0.009)	(0.017)	(0.014)
Actual wassce score			0.009***	0.009***			0.007***	0.007***
			(0.001)	(0.001)			(0.001)	(0.001)
Prcvd prob. admitted uni			0.001	0.001			0.001	0.001
			(0.001)	(0.001)			(0.001)	(0.001)
Prcvd prob. afford uni			0.002***	0.002***			0.001	0.001*
			(0.001)	(0.001)			(0.001)	(0.001)
Age	-0.030***	-0.033***	-0.015*	-0.017**	-0.023***	-0.024***	-0.015**	-0.016**
	(0.005)	(0.005)	(0.008)	(0.008)	(0.004)	(0.004)	(0.007)	(0.006)
Male	0.143***	0.140***	0.123***	0.112***	0.091***	0.089***	0.085***	0.068**
	(0.017)	(0.024)	(0.026)	(0.033)	(0.014)	(0.020)	(0.024)	(0.027)
Father primary educ	-0.029	-0.039	-0.016	-0.067	-0.020	-0.032	0.005	-0.023
	(0.045)	(0.044)	(0.071)	(0.069)	(0.039)	(0.033)	(0.064)	(0.055)
Father JHS educ	-0.037	-0.029	-0.093**	-0.113**	-0.033	-0.024	-0.068	-0.072*
	(0.027)	(0.029)	(0.046)	(0.049)	(0.024)	(0.022)	(0.041)	(0.041)
Father SHS educ	-0.018	-0.018	-0.051	-0.075	-0.019	-0.010	-0.037	-0.030
	(0.030)	(0.031)	(0.048)	(0.048)	(0.026)	(0.026)	(0.043)	(0.044)
Father post-sec educ	0.076**	0.073**	-0.021	-0.034	0.045*	0.041	-0.002	-0.015
	(0.031)	(0.031)	(0.051)	(0.048)	(0.027)	(0.027)	(0.046)	(0.045)
Mother primary educ	0.052*	0.060**	0.006	0.027	0.031	0.030	0.002	0.009
	(0.031)	(0.030)	(0.050)	(0.048)	(0.027)	(0.026)	(0.045)	(0.045)
Mother JHS educ	0.008	0.022	-0.011	0.024	0.014	0.019	-0.003	0.010
	(0.023)	(0.023)	(0.039)	(0.042)	(0.020)	(0.019)	(0.035)	(0.032)
Mother SHS educ	0.008	0.033	-0.035	-0.009	0.044*	0.054*	0.018	0.030
	(0.030)	(0.031)	(0.047)	(0.047)	(0.026)	(0.028)	(0.042)	(0.042)
Mother post-sec educ	0.097**	0.103**	0.028	0.035	0.045	0.049	-0.034	-0.024
	(0.040)	(0.042)	(0.061)	(0.061)	(0.034)	(0.032)	(0.055)	(0.045)
Number of siblings	-0.011***	-0.008**	-0.011**	-0.009**	-0.007**	-0.005*	-0.006	-0.004
	(0.004)	(0.004)	(0.005)	(0.004)	(0.003)	(0.003)	(0.005)	(0.004)
Born in rural area	-0.024	-0.016	-0.027	-0.015	-0.009	-0.007	-0.021	-0.018
	(0.017)	(0.018)	(0.026)	(0.027)	(0.015)	(0.016)	(0.023)	(0.025)
Inverse of Mill's ratio	-0.061***		-0.069***		-0.060***		-0.057***	
	(0.011)		(0.017)		(0.010)		(0.016)	
Constant	0.825***	0.747***	-0.142	-0.196	0.484***	0.400***	-0.150	-0.239
	(0.142)	(0.139)	(0.226)	(0.236)	(0.123)	(0.110)	(0.204)	(0.214)
Region fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3064	3314	1267	1323	3064	3314	1267	1323
R-squared	0.107	0.091	0.219	0.198	0.089	0.076	0.192	0.171

Note : *** p<0.01, ** p<0.05, * p<0.1. Clustered at SHS level standard errors are in parentheses. Anne Duplantier (2021)

TABLEAU 4.25 Corrected Estimation by gender - Heckman

	Application				Attendance			
	Men		Women		Men		Women	
	Corrected	Chap3	Corrected	Chap3	Corrected	Chap3	Corrected	Chap3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Exp. uncondi. returns	-6.53e-06	-0.016	-3.80e-06	0.001	-2.82e-06	-0.007	1.40e-06	-0.011
	(7.60e-06)	(0.012)	(1.09e-05)	(0.014)	(7.09e-06)	(0.012)	(9.55e-06)	(0.014)
Exp. max-min	0.012	0.013	0.037	0.042*	0.027	0.029	0.021	0.038*
	(0.025)	(0.021)	(0.028)	(0.021)	(0.023)	(0.019)	(0.025)	(0.021)
Actual wassce score	0.010***	0.011***	0.007***	0.007***	0.009***	0.009***	0.005***	0.005***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Prcvd prob. admitted uni	0.001	0.001	0.000	0.001	0.000	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Prcvd prob. afford uni	0.001	0.001*	0.002**	0.002***	0.001	0.001	0.001	0.001*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Age	-0.012	-0.015*	-0.021	-0.014	-0.012	-0.013*	-0.027**	-0.019
	(0.010)	(0.009)	(0.015)	(0.013)	(0.009)	(0.008)	(0.013)	(0.011)
Father primary educ	0.045	0.004	-0.187	-0.210*	0.048	0.006	-0.097	-0.083
	(0.089)	(0.092)	(0.121)	(0.125)	(0.083)	(0.069)	(0.105)	(0.095)
Father JHS educ	-0.141**	-0.125*	-0.070	-0.102	-0.115**	-0.111**	-0.046	-0.025
	(0.060)	(0.064)	(0.072)	(0.087)	(0.056)	(0.053)	(0.063)	(0.070)
Father SHS educ	-0.046	-0.058	-0.096	-0.108	-0.044	-0.032	-0.073	-0.042
	(0.062)	(0.060)	(0.075)	(0.085)	(0.058)	(0.050)	(0.066)	(0.081)
Father post-sec educ	0.022	0.043	-0.105	-0.127	0.001	-0.012	-0.041	-0.020
	(0.070)	(0.068)	(0.078)	(0.083)	(0.065)	(0.062)	(0.068)	(0.074)
Mother primary educ	-0.027	0.012	0.074	0.060	-0.007	0.005	0.036	0.019
	(0.066)	(0.068)	(0.077)	(0.073)	(0.061)	(0.058)	(0.067)	(0.063)
Mother JHS educ	0.040	0.036	0.036	0.032	0.046	0.029	0.009	-0.001
	(0.050)	(0.052)	(0.062)	(0.066)	(0.046)	(0.041)	(0.054)	(0.048)
Mother SHS educ	-0.013	-0.005	0.034	0.013	0.063	0.054	0.032	0.019
	(0.062)	(0.068)	(0.072)	(0.062)	(0.057)	(0.059)	(0.063)	(0.058)
Mother post-sec educ	0.031	0.033	0.099	0.043	0.009	0.000	-0.027	-0.049
	(0.084)	(0.090)	(0.089)	(0.080)	(0.078)	(0.066)	(0.077)	(0.069)
Number of siblings	-0.011	-0.009	-0.003	-0.005	-0.006	-0.005	-0.002	-0.003
	(0.007)	(0.007)	(0.008)	(0.006)	(0.007)	(0.007)	(0.007)	(0.006)
Born in rural area	-0.004	-0.000	-0.041	-0.024	-0.009	-0.004	-0.023	-0.030
	(0.034)	(0.032)	(0.040)	(0.038)	(0.032)	(0.030)	(0.035)	(0.038)
Inverse of Mill's ratio	-0.082***		-0.062**		-0.066***		-0.047*	
	(0.016)		(0.016)		(0.015)		(0.015)	
Constant	-0.156	-0.072	-0.176	-0.364	-0.279	-0.257	0.053	-0.188
	(0.291)	(0.286)	(0.376)	(0.336)	(0.271)	(0.278)	(0.328)	(0.305)
Region fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	755	803	512	520	755	803	512	520
R-squared	0.248	0.208	0.203	0.191	0.214	0.186	0.159	0.162

Note : *** p<0.01, ** p<0.05, * p<0.1. Clustered at SHS level standard errors are in parentheses. Anne Duplantier (2021)

4.8 CONCLUSION

As this dissertation uses a panel database, conducting an attrition analysis is necessary to know the potential bias and threat. This chapter presented the literature on attrition as well as the theoretical framework. It then described the situation of attrition in the sample. Four types of individuals are considered in order to give the most rigorous description of attrition in the sample. Over the 7616 individuals in the sample, 14% have never been interviewed and 47% have been interviewed as often as they should have been. The wave-attritors represent 12% of the sample, while respondents whole leave and re-enter the survey at different waves constitute 27% of the respondents. However, there are some differences according to the gender and the cohort, as well by region.

Following the literature, an attrition analysis is conducted to estimate the pattern and determinants of attrition. First, I looked at whether attritors differ from non-attritors regarding their characteristics. The results suggest that attrition is patterned. Attritors are on average younger and have a higher probability to come from rural areas and to have more siblings. Moreover, attritors are more likely to have less educated parents and a lower score at JHS.

Second, I used FGM and BLGW tests to estimate whether results from Chapter 3 suffer from an attrition bias. Some characteristics of the database do not allow us to analyze attrition bias directly and perfectly. That is why a substitution sample is built and used to run attrition tests. Results suggest that, in general, both education outcomes - application to and attendance at university - suffer from an attrition bias. It seems that attrition is significantly and negatively related to probabilities of application to and attendance at university. However, most of the attrition seems to be random or explained by unobservable characteristics.

Finally, a correction is applied to results reported in Chapter 3 in order to account for attrition. Results are very close to the ones found without correction. The size of coefficients changes slightly and is in general higher after correction. This suggests that the results in Chapter 3 are quite robust and consistent. Without correction, results are probably slightly underestimated but still significant. In conclusion, the database used in the thesis suffers from attrition, but the results reported in Chapter 3 are still valid.

CONCLUSION GÉNÉRALE

L'éducation et l'insertion des jeunes sur le marché du travail restent des enjeux primordiaux pour les pays en développement tel que le Ghana. En effet, les personnes les plus éduquées participent particulièrement à la croissance et la réduction de la pauvreté, grâce à leurs salaires plus élevés mais aussi à leur instruction les poussant à prendre soin de leur santé et de l'environnement. D'ailleurs les rendements de l'éducation supérieure sont les plus élevés du monde. Si des progrès sont notables en matière d'éducation primaire et secondaire, l'enseignement supérieur n'est pas encore très développé en Afrique Sub-Saharienne. D'ailleurs, l'éducation tertiaire dans les pays en développement est en général peu étudiée par les économistes.

Dans ce contexte, cette thèse a pour objectif d'apporter de l'information sur la situation des jeunes diplômés de l'école secondaire au Ghana qui poursuivent les études et/ou entrent sur le marché du travail. En particulier, deux principales questions sont abordées. Tout d'abord nous cherchons à comprendre comment les jeunes éduqués se déplacent au sein du Ghana pour avoir de meilleurs salaires. La deuxième question est de savoir comment ces jeunes prennent la décision de continuer leurs études au niveau supérieur après avoir diplômé de l'école secondaire. Cette thèse contribue donc en abordant des thèmes cruciaux du développement tels que l'éducation et le marché du travail et les mêlant à des concepts peu étudiés tels que la migration interne, les études supérieures, et les attentes dans les pays en développement. De plus, ce travail contribue en apportant une base de données primaire d'un échantillon d'une taille conséquente dans un pays en développement.

Le premier chapitre permet de présenter la base de données collectée et utilisée dans la thèse ainsi que les choix méthodologiques nécessaires pour la collecte, l'expérience de terrain et la gestion de la base de données. De plus, ce chapitre apporte une description détaillée et approfondie de l'éducation secondaire et post-secondaire au Ghana ainsi que la situation des jeunes éduqués sur le marché du travail. De plus, une attention particulière est apportée aux différences selon le genre afin de comprendre si les résultats d'éducation et de marché du travail varient selon que l'on est une fille ou un garçon.

Cette analyse descriptive permet de mettre en évidence plusieurs faits stylisés en matière d'éducation et de marché du travail. Les filles ont moins accès à l'école secondaire que les garçons, mais elles ont une plus grande probabilité de fréquenter une institution post-secondaire que les garçons. Il y a aussi des différences selon le genre dans le type d'institution fréquentée. Concernant les résultats touchant au marché du travail, les filles ont une plus forte probabilité d'être au chômage après leurs études. Les jeunes femmes éduquées ont donc plus de difficultés à trouver un emploi que leurs homologues masculins. En moyenne, les individus qui ont fréquenté une institution post-

secondaire ont des revenus plus élevés que les autres. Il y a donc une prime à l'éducation après le secondaire.

Le deuxième chapitre analyse les liens entre la migration interne des jeunes Ghanéens éduqués et le marché du travail. L'objectif est d'étudier le rôle des revenus espérés moyens régionaux et de leur variabilité dans la décision de changer de région de résidence des jeunes qui sortent de l'école secondaire. La méthodologie utilisée est un modèle de choix comportemental estimé par un logit mixte. Cette méthode permet de dépasser certaines limites du modèle logit standard telle que l'hypothèse d'indépendance des alternatives non pertinentes. Ce chapitre contribue à la littérature sur la migration interne dans les pays d'Afrique de l'Ouest en considérant la migration comme une décision individuelle et en choisissant comme population d'étude les jeunes éduqués. En effet, cette population est d'une grande importance pour le développement d'un pays mais souvent peu étudiée par manque de données apropriées. De plus, au lien de se concentrer uniquement sur les différences de revenus moyens comme déterminant de la migration, le risque est pris en compte en intégrant les différences de variations de revenus entre les régions.

Les résultats suggèrent qu'un revenu moyen suppérieur dans une autre région augmente la probabilité que les jeunes éduqués se déplacent dans cette région. Les différences régionales dans la variabilité des revenus sont aussi positivement liées à la migration interne. Ceci suggère que les jeunes éduqués migrants sont attirés par le risque ou bien qu'ils pensent pouvoir réussir mieux que les autres sur le marché du travail et être positionnés dans la fourchette haute des revenus. Enfin, les jeunes sortant de l'école secondaire qui sont nés dans une région rurale ont une plus faible probabilité de migrer, alors que la probabilité de migrer est positivement liée à l'éducation de la mère et aux capacités individuelles du jeune.

Le chapitre 3 porte sur le rôle des attentes dans les choix d'éducation post-secondaire au Ghana. L'objectif est d'étudier le lien entre les rendements espérés de l'éducation et les décisions de postuler et de fréquenter l'université. De plus, le rôle des capacités espérées et réelles est analysé. La méthodologie utilisée est un modèle linéaire de choix. Ce travail contribue à la littérature en décomposant le processus de choix d'éducation post-secondaire en deux étapes : d'une part la candidature et de l'autre la fréquentation. De plus, au lieu de se concentrer uniquement sur les rendements espérés de l'éducation comme déterminant des choix d'éducation, deux nouvelles composantes sont ajoutées : les capacités réelles de l'individu, mesurées par les résultats académiques secondaires, ainsi que les capacités perçues, mesurées avant que l'individu connaisse ses résultats académiques.

Les résultats montrent que les rendements espérés de l'éducation sont positivement liés à la décision de candidater à l'université. Cependant, une fois que l'on prend en compte les capacités réelles et estimées, les rendements espérés ne sont plus significativement liés aux décisions d'éducation post-secondaire. Les capacités réelles et perçues sont positivement liées à la probabilité de postuler

à l'université ainsi que la décision de fréquenter l'université. Il est donc nécessaire et important de prendre en compte les capacités quand on étudie la décision de s'éduquer.

Le quatrième et dernier chapitre aborde la question technique du traitement de l'attrition qui touche la base de données en panel utilisée dans cette thèse. L'objectif est de conduire une analyse d'attrition à partir de cette base de données de panel. Afin de comprendre l'ampleur et les charactéristiques de l'attrition, une première partie de ce chapitre porte sur la description de l'attrition ainsi que les caractéristiques des individus sujets à l'attrition. Dans un deuxième temps, les déterminants de l'attrition sont analysés. Enfin des tests sont mis en oeuvre afin de comprendre si les résultats d'intérêt des chapitres précédents souffrent d'un biais d'attrition, et le cas échéant, une correction est appliquée. Alors que la grande majorité des travaux sur l'attrition se concentrent sur l'attrition venant des caractéristiques observables, les spécificités de notre base de données nous permettent d'étudier également l'attrition venant des caractéristiques inobservables.

Les résultats suggèrent que les individus sujets à l'attrition ont des caractéristiques différentes des autres. Ces individus sont en moyenne plus jeunes, ont une plus grande probabilité de venir de zones rurales et ont en moyenne plus de frères et soeurs. De plus, les individus sujets à l'attrition ont en général des parents moins éduqués et de moins bons résultats scolaires. Les résultats des tests montrent que les résultats d'intérêt souffrent d'un biais d'attrition. En effet, les probabilités de candidater et de fréquenter l'université sont négativement liés à la probabilité d'être sujet à l'attrition. Après correction, les résultats sont très proches de ceux non corrigés. Même si la base de données est sujette à l'attrition, les résultats restent pertinents et valides.

Bibliographie

Ackah, C. and Medvedev, D. (2012). Internal migration in ghana : determinants and welfare impacts. *International Journal of Social Economics*, 39(10) :764–784.

Ackah Baidoo, P. (2016). Youth unemployment in resource-rich sub-saharan africa : A critical review. *The extractive industries and society*, 3(1) :249–261.

Adams, R. H. and Page, J. (2005). Do international migration and remittances reduce poverty in developing countries ? *World development*, 33(10) :1645–1669.

Agnew, J. R., Anderson, L. R., Gerlach, J. R., and Szykman, L. R. (2008). Who chooses annuities ? an experimental investigation of the role of gender, framing, and defaults. *The American Economic Review*, 98(2) :418–422.

Ajayi, K. F. (2014). Does school quality improve student performance ? new evidence from ghana. *Unpublished working paper*.

Akgüç, M., Liu, X., Tani, M., and Zimmermann, K. F. (2016). Risk attitudes and migration. *China Economic Review*, 37 :166–176.

Alderman, H., Behrman, J. R., Kohler, H.-P., Maluccio, J. A., and Watkins, S. C. (2001). Attrition in longitudinal household survey data : Some tests for three developing-country samples. *Demographic research*, 5 :79–124.

Álvarez, B. and Vera-Hernández, M. (2013). Exploiting subjective information to understand impoverished children's use of health care. *Journal of Health Economics*, 32(6) :1194–1204.

Anarfi, J., Kwankye, S., Ababio, O.-M., and Tiemoko, R. (2003). Migration from and to ghana : A background paper. *University of Sussex : DRC on Migration, Globalisation and Poverty*.

Angrist, J. D. (2001). Estimation of limited dependent variable models with dummy endogenous regressors : Simple strategies for empirical practice. *Journal of business & economic statistics*, 19(1) :2–28.

Arcidiacono, P., Hotz, V. J., and Kang, S. (2012). Modeling college major choices using elicited measures of expectations and counterfactuals. *Journal of Econometrics*, 166(1) :3–16.

Arrow, K. J. (1973). Information and economic behavior. Technical report, HARVARD UNIV CAMBRIDGE MA.

Ashraf, N. (2009). Spousal control and intra-household decision making : An experimental study in the philippines. *American Economic Review*, 99(4) :1245–77.

Assaad, R., Krafft, C., et al. (2016). Labor market dynamics and youth unemployment in the middle east and north africa : evidence from egypt, jordan and tunisia. In *Economic Research Forum Working Paper Series*, volume 993, pages 1–27.

Attanasio, O. and Kaufmann, K. (2009). Educational choices, subjective expectations, and credit constraints. Technical report, National Bureau of Economic Research.

Attanasio, O., Kugler, A., and Meghir, C. (2007). Effects of youth training in developing countries : Evidence from a randomized training program in colombia. *Londres : University College London*.

Attanasio, O. P. (2009). Expectations and perceptions in developing countries : Their measurement and their use. *American Economic Review*, 99(2) :87–92.

Attanasio, O. P. and Kaufmann, K. M. (2014). Education choices and returns to schooling : Mothers and youths subjective expectations and their role by gender. *Journal of Development Economics*, 109 :203–216.

Attanasio, O. P. and Kaufmann, K. M. (2017). Education choices and returns on the labor and marriage markets : Evidence from data on subjective expectations. *Journal of Economic Behavior & Organization*, 140 :35–55.

Avery, C. and Kane, T. J. (2004). Student perceptions of college opportunities. the boston coach program. In *College choices : The economics of where to go, when to go, and how to pay for it*, pages 355–394. University of Chicago Press.

Avura, F. B. and Ulzen-Appiah, A. (2016). Ghana youth employment program inventory. *The World Bank Working Paper*.

Awumbila, M., Manuh, T., Quartey, P., Tagoe, C. A., and Bosiakoh, T. A. (2008). Migration country paper (ghana). *Centre for Migration Studies, University of Ghana, Legon*.

Baah-Boateng, W. (2013). Determinants of unemployment in ghana. *African Development Review*, 25(4) :385–399.

Baigrie, N. and Eyal, K. (2014). An evaluation of the determinants and implications of panel attrition in the national income dynamics survey (2008-2010). *South African Journal of Economics*, 82(1) :39–65.

Barsky, R. B., Juster, F. T., Kimball, M. S., and Shapiro, M. D. (1997). Preference parameters and behavioral heterogeneity : An experimental approach in the health and retirement study. *The Quarterly Journal of Economics*, 112(2) :537–579.

Bashir, S., Lockheed, M., Ninan, E., and Tan, J.-P. (2018). Facing forward : Schooling for learning in africa. Africa development forum series, World Bank.

Baulch, B. and Quisumbing, A. (2011). Testing and adjusting for attrition in household panel data. *CPRC Toolkit Note*.

Becker, G. S. (1964). Human capital theory. *Columbia University Press, New York*.

Becketti, S., Gould, W., Lillard, L., and Welch, F. (1988). The panel study of income dynamics after fourteen years : An evaluation. *Journal of Labor Economics*, 6(4) :472–492.

Bertoli, S. and Marchetta, F. (2014). Migration, remittances and poverty in ecuador. *The Journal of Development Studies*, 50(8) :1067–1089.

Betts, J. R. (1996). What do students know about wages ? evidence from a survey of undergraduates. *Journal of Human Resources*, pages 27–56.

Binswanger, H. P. (1980). Attitudes toward risk : Experimental measurement in rural india. *American journal of agricultural economics*, 62(3) :395–407.

Binswanger, H. P. (1981). Attitudes toward risk : Theoretical implications of an experiment in rural india. *The Economic Journal*, 91(364) :867–890.

Bonin, H., Constant, A., Tatsiramos, K., and Zimmermann, K. F. (2009). Native-migrant differences in risk attitudes. *Applied Economics Letters*, 16(15) :1581–1586.

Borghans, L., Heckman, J. J., Golsteyn, B. H., and Meijers, H. (2009). Gender differences in risk aversion and ambiguity aversion. *Journal of the European Economic Association*, 7(2-3) :649–658.

Boyd, J. H. and Mellman, R. E. (1980). The effect of fuel economy standards on the us automotive market : an hedonic demand analysis. *Transportation Research Part A : General*, 14(5-6) :367–378.

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources*, 50(2) :317–372.

Cameron, A. C. and Trivedi, P. K. (2010). *Microeconometrics using stata*, volume 2. Stata press College Station, TX.

Cardell, N. S. and Dunbar, F. C. (1980). Measuring the societal impacts of automobile downsizing. *Transportation Research Part A : General*, 14(5-6) :423–434.

Chiswick, B. R. (1978). The effect of americanization on the earnings of foreign-born men. *Journal of political Economy*, 86(5) :897–921.

Cho, Y., Robalino, D., and Watson, S. (2016). Supporting self-employment and small-scale entrepreneurship : potential programs to improve livelihoods for vulnerable workers. *IZA Journal of Labor Policy*, 5(1) :7.

Constant, A. and Zimmermann, K. F. (2006). The making of entrepreneurs in germany : Are native men and immigrants alike ? *Small business economics*, 26(3) :279–300.

Darvas, P., Gao, S., Shen, Y., and Bawany, B. (2017). Sharing higher education's promise beyond the few in sub-saharan africa. Direction in development, World Bank.

De Vreyer, P., Gubert, F., and Roubaud, F. (2009). Migration, self-selection and returns to education in the waemu. *Journal of African Economies*, 19(1) :52–87.

Delavande, A., Giné, X., and McKenzie, D. (2011). Measuring subjective expectations in developing countries : A critical review and new evidence. *Journal of Development Economics*, 94(2) :151–163.

Delavande, A. and Kohler, H.-P. (2009). Subjective expectations in the context of hiv/aids in malawi. *Demographic research*, 20 :817.

Deshingkar, P. and Grimm, S. (2005). *Internal migration and development : a global perspective*. Number 19. United Nations Publications.

Ding, X., Hartog, J., and Sun, Y. (2010). Can we measure individual risk attitudes in a survey ?

Dominitz, J. and Manski, C. F. (1996). Eliciting student expectations of the returns to schooling. Technical Report 1.

Donkers, B., Melenberg, B., van Soest, A., et al. (1999). *Estimating risk attitudes using lotteries : a large sample approach*. Tilburg University.

Du, Y., Park, A., and Wang, S. (2005). Migration and rural poverty in china. *Journal of comparative economics*, 33(4) :688–709.

Dudwick, N. (2011). *From farm to firm : Rural-urban transition in developing countries*. World Bank publications.

Duflo, E., Dupas, P., and Kremer, M. (2017). The impact of free secondary education : Experimental evidence from ghana. *Massachusetts Institute of Technology Working Paper Cambridge, MA*.

Duflo, E. and Saez, E. (2003). The role of information and social interactions in retirement plan decisions : Evidence from a randomized experiment. *The Quarterly Journal of Economics*, 118(3) :815–842.

Dupas, P. (2011). Do teenagers respond to hiv risk information ? evidence from a field experiment in kenya. *American Economic Journal : Applied Economics*, 3(1) :1–34.

Eckel, C. C. and Grossman, P. J. (2008). Men, women and risk aversion : Experimental evidence. *Handbook of experimental economics results*, 1 :1061–1073.

EP-Nuffic (2015). The ghanaian education system described and compared with the dutch system. Technical report.

Fafchamps, M. and Shilpi, F. (2013). Determinants of the choice of migration destination. *Oxford Bulletin of Economics and Statistics*, 75(3) :388–409.

Falaris, E. M. (2003). The effect of survey attrition in longitudinal surveys : evidence from peru, côte d'ivoire and vietnam. *Journal of Development Economics*, 70(1) :133–157.

Filmer, D. and Fox, L. (2014). Youth employment in sub-saharan africa. *Africa Development Series*.

Fitzgerald, J., Gottschalk, P., and Moffitt, R. A. (1998). An analysis of sample attrition in panel data : The michigan panel study of income dynamics.

Fox, L., Senbet, L. W., and Simbanegavi, W. (2016). Youth employment in sub-saharan africa : challenges, constraints and opportunities. *Journal of African Economies*, 25(suppl_1) :i3–i15.

Friedman, J. (2012). Whether to probit or to probe it : In defense of the linear probability model. https://blogs.worldbank.org/impactevaluations/whether-to-probit-or-to-probe-it-in-defense-of-the-linear-probability-mode Accessed : 2019-06-11.

Fuwa, N. (2011). Should we track migrant households when collecting household panel data ? household relocation, economic mobility, and attrition biases in the rural philippines. *American Journal of Agricultural Economics*, 93(1) :56–82.

Ghana Statistical Service (2012). Summary report of final results from the 2010 population and housing census. Technical report, Ghana Statistical Service.

Ghana Statistical Service (2014a). Ghana living standards survey round 6 (glss6) - main report. Technical report.

Ghana Statistical Service (2014b). Poverty profile in ghana (2005-2013). Technical report.

Ghana Statistical Service (2019). Ghana statistical service website. https://statsghana.gov.gh/. Accessed : 2019-11-25.

Ghana Statistical Service, Ghana Health Service, and ICF International (2015). Ghana demographic and health survey 2014. Technical report, Ghana Statistical Service.

Gibson, J. and McKenzie, D. (2011). The microeconomic determinants of emigration and return migration of the best and brightest : Evidence from the pacific. *Journal of Development Economics*, 95 :18–29.

Goldbach, C. and Schlüter, A. (2018). Risk aversion, time preferences, and out-migration. experimental evidence from ghana and indonesia. *Journal of Economic Behavior & Organization*, 150 :132–148.

Gomila, R. (2019). Logistic or linear ? estimating causal effects of treatments on binary outcomes using regression analysis.

Gray, C. L. (2009). Environment, land, and rural out-migration in the southern ecuadorian andes. *World Development*, 37(2) :457–468.

Griliches, Z. (1977). Estimating the returns to schooling : Some econometric problems. *Econometrica : Journal of the Econometric Society*, pages 1–22.

Guiso, L. and Paiella, M. (2004). The role of risk aversion in predicting individual behaviors.

Hao, L., Houser, D., Mao, L., and Villeval, M. C. (2016). Migrations, risks, and uncertainty : A field experiment in china. *Journal of Economic Behavior & Organization*, 131 :126–140.

Hare, D. (1999). 'push' versus 'pull' factors in migration outflows and returns : Determinants of migration status and spell duration among china's rural population. *The Journal of Development Studies*, 35(3) :45–72.

Harris, J. R. and Todaro, M. P. (1970). Migration, unemployment and development : a two-sector analysis. *The American economic review*, 60(1) :126–142.

Harrison, G. W., Lau, M. I., and Rutström, E. E. (2007). Estimating risk attitudes in denmark : A field experiment. *The Scandinavian Journal of Economics*, 109(2) :341–368.

Hartog, J., Ferrer-i Carbonell, A., and Jonker, N. (2002). Linking measured risk aversion to individual characteristics. *Kyklos*, 55(1) :3–26.

Hastings, J. S. and Weinstein, J. M. (2008). Information, school choice, and academic achievement : Evidence from two experiments. *The Quarterly journal of economics*, 123(4) :1373–1414.

Heckman, J. (1979). Sample specification bias as a selection error. *Econometrica*, 47(1) :153–162.

Heitmueller, A. (2005). Unemployment benefits, risk aversion, and migration incentives. *Journal of Population Economics*, 18(1) :93–112.

Hensher, D. A. and Greene, W. H. (2003). The mixed logit model : the state of practice. *Transportation*, 30(2) :133–176.

Hicks, N. (1980). *Economic growth and human resources*. World Bank.

Hill, D. H. and Willis, R. J. (2001). Reducing panel attrition : A search for effective policy instruments. *Journal of Human Resources*, pages 416–438.

Hippel, P. V. (2015). Linear versus logistic probability models : Which is better, and when ? `https://statisticalhorizons.com/linear-vs-logistic`. Accessed : 2019-06-12.

Hogan, V. and Rigobon, R. (2002). Using heteroscedasticity to estimate the returns to education. *National Bureau of Economic Research Working Paper*.

Hoogeveen, J., Croke, K., Dabalen, A., Demombynes, G., and Giugale, M. (2014). Collecting high frequency panel data in africa using mobile phone interviews. *Canadian Journal of Development Studies/Revue canadienne d'études du développement*, 35(1) :186–207.

Horrace, W. C. and Oaxaca, R. L. (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, 90(3) :321–327.

Hurd, M. D. (2009). Subjective probabilities in household surveys. *Annu. Rev. Econ.*, 1(1) :543–562.

Jaeger, D. A., Dohmen, T., Falk, A., Huffman, D., Sunde, U., and Bonin, H. (2010). Direct evidence on risk attitudes and migration. *The Review of Economics and Statistics*, 92(3) :684–689.

Jensen, R. (2010). The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics*, 125(2) :515–548.

Kalton, G. and Brick, M. (2000). Weighting in household panel surveys. *in Rose, D. (ed) Researching social and economic change : the uses of household panel studies*, pages 94–107.

Katz, E. and Stark, O. (1986). Labor migration and risk aversion in less developed countries. *Journal of labor Economics*, 4(1) :134–149.

Kaufmann, K. M. (2014). Understanding the income gradient in college attendance in mexico : The role of heterogeneity in expected returns. *Quantitative Economics*, 5(3) :583–630.

Lall, S. V., Timmins, C., and Yu, S. (2009). Connecting lagging and leading regions : The role of labor mobility. *Brookings-Wharton Papers on Urban Affairs*, 2009(1) :151–174.

Laurie, H. and Lynn, P. (2008). *The use of respondent incentives on longitudinal surveys*, volume 42. University of Essex, Institute for Social and Economic Research (ISER), Colchester.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, 48(2) :281–355.

Lee, U. (2003). Panel attrition in survey data : a literature review. *University of Cape Town Working Paper*, 41.

Lewbel, A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics*, 30(1) :67–80.

Lewis, W. A. (1954). Economic development with unlimited supplies of labour. *The manchester school*, 22(2) :139–191.

Lillard, L. A. and Panis, C. W. (1998). Panel attrition from the panel study of income dynamics : Household income, marital status, and mortality. *Journal of Human Resources*, pages 437–457.

Lucas, R. E. (1997). Internal migration in developing countries. *Handbook of population and family economics*, 1 :721–798.

Magruder, J. and Nattrass, N. (2006). Exploring attrition bias : the case of the khayelitsha panel study (2000-2004). *South African Journal of Economics*, 74(4) :769–781.

Maluccio, J. et al. (2000). Attrition in the kwazulu natal income dynamics study, 1993-1998. *Food Consumption and Nutrition Division Discussion Paper*, 95.

Maluccio, J. A. (2004). Using quality of interview information to assess nonrandom attrition bias in developing-country panel data. *Review of Development Economics*, 8(1) :91–109.

Manski, C. F. (2004). Measuring expectations. *Econometrica*, 72(5) :1329–1376.

Mayda, A. M. (2010). International migration : A panel data analysis of the determinants of bilateral flows. *Journal of Population Economics*, 23(4) :1249–1274.

McFadden, D. and Train, K. (2000). Mixed mnl models for discrete response. *Journal of applied Econometrics*, pages 447–470.

McFadden, D. L. (1974). Conditional logit analysis of qualitative choice behavior. *Chapter 4 in P. Zaremba, ed. Frontiers in Econometrics*.

McKenzie, D., Gibson, J., and Stillman, S. (2013). A land of milk and honey with streets paved with gold : Do emigrants have over-optimistic expectations about incomes abroad ? *Journal of Development Economics*, 102 :116–127.

McKenzie, D. J. and Sasin, M. J. (2007). Migration, remittances, poverty, and human capital : conceptual and empirical challenges. *World Bank Policy Research Working Paper 4272*.

Miguel, E. and Hamory, J. (2009). Individual ability and selection into migration in kenya. *Human Development Reports Research Paper*.

Ministry of Education Ghana (2018a). Education sector analysis. Technical report, Ministry of Education Ghana.

Ministry of Education Ghana (2018b). Education sector medium-term development plan 2018-2021. Technical report, Ministry of Education Ghana.

Ministry of Education Ghana (2018c). Education strategic plan 2018-2030. Technical report, Ministry of Education Ghana.

Mora, J. and Taylor, J. E. (2006). Determinants of migration, destination, and sector choice : Disentangling individual, household, and community effects. *International migration, remittances, and the brain drain*, pages 21–52.

Nations Unies (2015). Objectifs du millénaire pour le développement. rapport 2015. Technical report.

Nelson, J. A. (2015). Are women really more risk-averse than men ? a re-analysis of the literature using expanded methods. *Journal of Economic Surveys*, 29(3) :566–585.

Nguyen, T. (2008). Information, role models and perceived returns to education : Experimental evidence from madagascar. *Unpublished manuscript*, 6.

Oberg, E. (2012). Ghana education sector : background paper.

Osei-Boateng, C. and Ampratwum, E. (2011). *The informal sector in Ghana*. Friedrich-Ebert-Stiftung, Ghana Office Accra.

Outes-Leon, I. and Dercon, S. (2009). Survey attrition and attrition bias in young lives. *Young Lives Technical Note*, 5.

Pitan, O. S. and Adedeji, S. (2012). Skills mismatch among university graduates in the nigeria labor market. *US-China Education Review*.

Poku-Boansi, M. and Afrane, S. (2016). Magnitude and impact of youth unemployment in ghana.

Rogers, T. F. (1976). Interviews by telephone and in person : Quality of responses and field performance. *Public Opinion Quarterly*, 40(1) :51–65.

Rouse, C. E. (2004). Low-income students and college attendance : An exploration of income expectations. *Social Science Quarterly*, 85(5) :1299–1317.

Ryu, E., Couper, M. P., and Marans, R. W. (2005). Survey incentives : Cash vs. in-kind ; face-to-face vs. mail ; response rate vs. nonresponse error. *International Journal of Public Opinion Research*, 18(1) :89–106.

Sackey, H. A. and Osei, B. (2006). Human resource underutilization in an era of poverty reduction : An analysis of unemployment and underemployment in ghana. *African development review*, 18(2) :221–247.

Schultz, T. W. (1961). Investment in human capital. *The American Economic Review*, pages 1–17.

Schwartz, A. (1976). Migration, age, and education. *Journal of Political Economy*, 84(4, Part 1) :701–719.

Schweri, J. and Hartog, J. (2017). Do wage expectations predict college enrollment ? evidence from healthcare. *Journal of Economic Behavior & Organization*, 141 :135–150.

Sequeira, S., Spinnewijn, J., and Xu, G. (2016). Rewarding schooling success and perceived returns to education : Evidence from india. *Journal of Economic Behavior & Organization*, 131 :373–392.

Shuy, R. W. (2003). *In-person versus telephone interviewing*. Sage Publications.

Singer, E. et al. (2002). The use of incentives to reduce nonresponse in household surveys. *Survey nonresponse*, 51 :163–177.

Singer, E., Van Hoewyk, J., Gebler, N., and McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics*, 15(2) :217.

Smith, H. L. and Powell, B. (1990). Great expectations : Variations in income expectations among college seniors. *Sociology of Education*, pages 194–207.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3) :355–374.

Stark, O. (1981). On the optimal choice of capital intensity in ldcs with migration. *Journal of Development Economics*, 9(1) :31–41.

Stark, O. and Bloom, D. E. (1985). The new economics of labor migration. *The american Economic review*, 75(2) :173–178.

Stinebrickner, R. and Stinebrickner, T. R. (2013). A major in science ? initial beliefs and final outcomes for college major and dropout. *Review of Economic Studies*, 81(1) :426–472.

Sturges, J. E. and Hanrahan, K. J. (2004). Comparing telephone and face-to-face qualitative interviewing : a research note. *Qualitative research*, 4(1) :107–118.

Sutherland-Addy, E. (2008). *Gender equity in junior and senior secondary education in sub-Saharan Africa*. The World Bank.

Thomas, D., Frankenberg, E., and Smith, J. P. (2001). Lost but not forgotten : Attrition and follow-up in the indonesia family life survey. *Journal of Human resources*, pages 556–592.

Tiongson, E. R. and Fares, J. (2007). *Youth unemployment, labor market transitions, and scarring : evidence from Bosnia and Herzegovina, 2001-04*. World Bank.

Todaro, M. (1980). Internal migration in developing countries : a survey. In *Population and economic change in developing countries*, pages 361–402. University of Chicago Press.

Todaro, M. P. (1969). A model of labor migration and urban unemployment in less developed countries. *The American economic review*, 59(1) :138–148.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Tsegai, D. (2007). Migration as a household decision : What are the roles of income differences ? insights from the volta basin of ghana. *The European journal of development research*, 19(2) :305–326.

United Nations Development Programme (2020). Latest human development index ranking. `http://hdr.undp.org/en/content/latest-human-development-index-ranking`. Accessed : 2021-01-11.

Vaillant, J. (2013). Attrition and follow-up rules in panel surveys : Insights from a tracking experience in madagascar. *Review of Income and Wealth*, 59(3) :509–538.

Watson, D. (2003). Sample attrition between waves 1 and 5 in the european community household panel. *European Sociological Review*, 19(4) :361–378.

Winkel, J. W. and Withers, S. D. (2000). Panel attrition. *Rose, D. (ed) Researching social and economic change : the uses of household panel studies*, pages 82–93.

Wiswall, M. and Zafar, B. (2014). Determinants of college major choice : Identification using an information experiment. *The Review of Economic Studies*, 82(2) :791–824.

Wooldridge, J. (2002). *Econometric analysis of cross section and panel data*. MIT press, Cambridge, Massachusetts.

World Bank (2019). Wdi website - ghana. `https://data.worldbank.org/country/ghana?view=chart`. Accessed : 2019-11-25.

World Bank (2020). Wdi website - ghana - school enrollment. `https://data.worldbank.org/indicator/SE.SEC.NENR?end=2019&name_desc=false&start=2019&view=bar`. Accessed : 2020-04-12.

WorldBank (2018). Learning to realize education's promise. World development report, The World Bank.

Zabel, J. E. (1998). An analysis of attrition in the panel study of income dynamics and the survey of income and program participation with an application to a model of labor market behavior. *Journal of Human Resources*, pages 479–506.

Zafar, B. (2011a). Can subjective expectations data be used in choice models ? evidence on cognitive biases. *Journal of Applied Econometrics*, 26(3) :520–544.

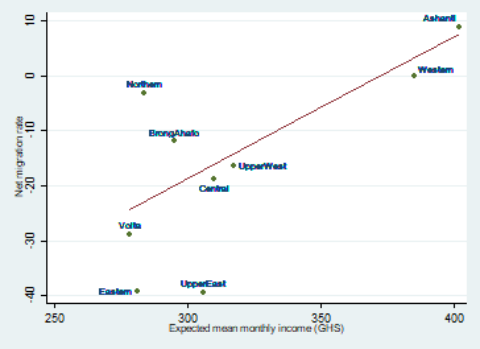
Zafar, B. (2011b). How do college students form expectations ? *Journal of Labor Economics*, 29(2) :301–348.

Zhang, K. H. and Shunfeng, S. (2003). Rural–urban migration and urbanization in china : Evidence from time-series and cross-section analyses. *China Economic Review*, 14(4) :386–400.

ANNEXE A

A.1 Expected Income and Migration

FIGURE A.1 Relationship Between Expected Average Income and Net Migration to Each Region - Without Greater Accra Region



Note : Relationship positive ($\beta = 0.698$) and statistically significant at 5% (p-value = 0.036). Anne Duplantier (2021)

ANNEXE B

B.1 The Probability Practice Questions

The questions of the probability practice are the following :

"1. Out of 10, how likely is it that it will rain tomorrow ?"
"2. During the rainy season, out of 10, how likely is it that it will rain on any given day ?"
"3. Think of another example. A woman who is pregnant but does not know if the baby is a boy or a girl. Out of 10, how likely is it that the baby will be a boy ?"
"4. Out of 10, how likely is it that you will go to the market tomorrow ?"
"5. Out of 10, how likely is it that you will get married over the next five years ?"
"6. Out of 10, how likely is it that you will get married over the next ten years ?"

B.2 The Triangular Distribution

The triangular distribution is composed of three parameters $a \leq c \leq b$ where a is the minimum, b is the maximum and c is the peak of the data. This kind of distribution is particularly useful in presence of limited sample data available. [1] These points are linearly connected to estimate the probability density function of the data. This later takes the following form :

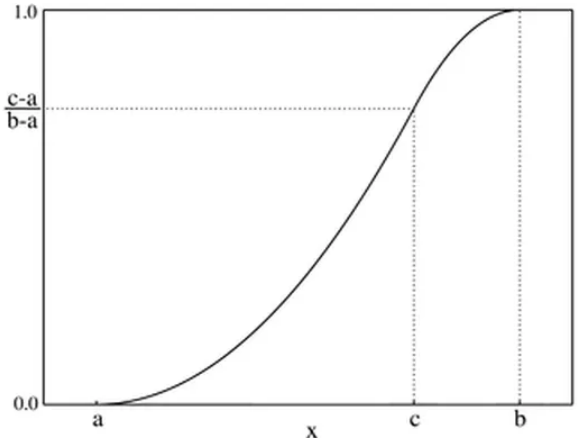
$$\begin{aligned} f(x|a,b,c) &= 0 && \text{if } x < a \\ &= \frac{2(x-a)}{(b-a)(c-a)} && \text{if } a \leq x \leq c \\ &= \frac{2(b-x)}{(b-a)(b-c)} && \text{if } c \leq x \leq b \\ &= 0 && \text{if } x > b \end{aligned} \tag{B.1}$$

And the cumulative density function (cf Figure B.1 for a visual representation) takes the form :

$$\begin{aligned} F(x|a,b,c) &= 0 && \text{if } x \leq a \\ &= \frac{(x-a)^2}{(b-a)(c-a)} && \text{if } a < x \leq c \\ &= 1 - \frac{(b-x)^2}{(b-a)(b-c)} && \text{if } c < x < b \\ &= 1 && \text{if } x \geq b \end{aligned} \tag{B.2}$$

1. cf https://www.mathworks.com/help/stats/triangular-distribution.html

FIGURE B.1 Triangular Cumulative Density Function



Source : MathWorks

ANNEXE C

C.1 Categories of Respondents - More Details

Table C.1 presents the categories of respondent, distinguished according to the theoretical number of waves. For wave-attritors and temporary stayers, statistics are detailed according to the number of waves in which they have really been interviewed. Among the respondents who should have been interviewed three times, 50.83% of them have really been interviewed three times (permanent stayers). 33.68% of them are total attritors and have never been interviewed. 8.13% of them are wave-attritors : 2.02% have been interviewed only one time and 6.11% two out of three times. Finally, the temporary stayers are 7.35% of this group and the majority of them have been interviewed two out of three times (but not successively as the wave-attritors). The other columns are presenting the same patterns but for the group of respondents who should have been interviewed four and five times. The more the theoretical number of waves increases, the more the likelihood to have temporary stayers is high (7.35% for the three-wave group, 23.86% for the four-wave group and 51.61% for the five-wave group).

TABLEAU C.1 Categories of Respondent by Number of Theoretical Waves

	Theoretical number of waves					
Type of		3 waves		4 waves		5 waves
respondent	Real/Theo	($N = 2176$)	Real/Theo	($N = 3260$)	Real/Theo	($N = 2176$)
Permanent stayers	3/3	1106	4/4	1821	5/5	682
		[50.83]		[55.86]		[31.34]
Total attritors	0/3	733	0/4	191	0/5	124
		[33.68]		[5.86]		[5.7]
All wave-attritors		177		474		247
		[8.13]		[14.54]		[11.35]
Detailed wave-attritors	1/3	44	1/4	254	1/5	139
		[2.02]		[7.79]		[6.39]
	2/3	133	2/4	92	2/5	51
		[6.11]		[2.82]		[2.34]
			3/4	128	3/5	15
				[3.93]		[0.69]
					4/5	42
						[1.93]
All temporary stayers		160		778		1123
		[7.35]		[23.86]		[51.61]
Detailed temp. stayers	1/3	9	1/4	54	1/5	28
		[0.41]		[1.66]		[1.29]
	2/3	151	2/4	205	2/5	87
		[6.94]		[6.29]		[4]
			3/4	519	3/5	292
				[15.92]		[13.42]
					4/5	716
						[32.9]

Note : One observation per respondent ($N = 7616$). Percentage is in brackets. Anne Duplantier (2021)

C.2 Definition of Variables

TABLEAU C.2 Variable Definition

Variable Name	Definition
Appl_uni	Dummy equals to 1 if the respondent applied to university
Att_uni	Dummy equals to 1 if the respondent attended university
Exp_return_uncondi	Difference between expected earnings of university and expected earnings of having no post-secondary education, multiplied by the probability of working
Exp_maxmin	Spread between the maximum and minimum expected earnings
Wasscescore	WASSCE score reversed and normalized from 0 to 100
Prcvd_prob_admitted_uni	Probability that the WASSCE results will be good enough to enter into university
Prcvd_prob_afford_uni	Probability to be able to afford attending university as a regular student if admitted
Age	Age of respondent in number of years
Gender	Dummy equals to 1 if the respondent is a man
Father_education_1	Dummy equals to 1 if father's respondent has no education
Father_education_2	Dummy equals to 1 if father's respondent highest education level is primary
Father_education_3	Dummy equals to 1 if father's respondent highest education level is JHS
Father_education_4	Dummy equals to 1 if father's respondent highest education level is SHS
Father_education_5	Dummy equals to 1 if father's respondent highest education level is post-secondary education
Mother_education_1	Dummy equals to 1 if mother's respondent has no education
Mother_education_2	Dummy equals to 1 if mother's respondent highest education level is primary
Mother_education_3	Dummy equals to 1 if mother's respondent highest education level is JHS
Mother_education_4	Dummy equals to 1 if mother's respondent highest education level is SHS
Mother_education_5	Dummy equals to 1 if mother's respondent highest education level is post-secondary education
Numsiblings	Number of siblings of the respondent
Ruralborn	Dummy equals 1 if the respondent is born in a rural area
Cohort_FE	Fixed effects at the cohort level : dummy variable for each cohort
Region_FE	Fixed effects at the region level : dummy variable for each region
Time_interview	Number of minutes the interview lasted
Participation_enum	Number of times the enumerator participated in a wave of the survey

C.3 Full Tables from Attrition Analysis

TABLEAU C.3 Tests of Wave Attrition - Substitution sample - OLS

	FGM test		BGLW test (Baseline value of)	
Dependent var.	Wave-attrition		Apply to uni	Attend at uni
	(1)	(2)	(3)	(4)
Applied to university	-0.037**			
	(0.014)			
Attended university		-0.045***		
		(0.012)		
Wave-attritors			-0.115	-0.555**
			(0.404)	(0.253)
Actual WASSCE score	0.001	0.001	-0.017***	-0.022***
	(0.001)	(0.001)	(0.002)	(0.002)
Attrit*WASSCE_score			0.000	0.005
			(0.004)	(0.003)
Exp. uncondi returns	0.009		0.005	
	(0.009)		(0.007)	
Attrit*return			0.000	
			(0.000)	
Exp. max-min	0.006		0.010	
	(0.012)		(0.016)	
Attrit*max_min			-0.032	
			(0.040)	
Age	-0.001	0.002	-0.019***	-0.016***
	(0.005)	(0.004)	(0.006)	(0.005)
Attrit*age			0.017	0.016*
			(0.012)	(0.009)
Male	-0.002	-0.003	0.131***	0.099***
	(0.016)	(0.012)	(0.028)	(0.022)
Attrit*male			-0.155**	-0.088*
			(0.067)	(0.050)
Father primary educ	0.053	0.060*	-0.047	-0.063
	(0.045)	(0.035)	(0.062)	(0.051)
Attrit*father_primary			-0.029	0.021
			(0.105)	(0.124)
Father JHS educ	-0.008	-0.011	-0.036	-0.050
	(0.025)	(0.019)	(0.042)	(0.034)
Attrit*father_JHS			0.198*	-0.014
			(0.115)	(0.083)
Father SHS educ	0.031	0.018	-0.024	-0.010
	(0.029)	(0.023)	(0.041)	(0.039)
Attrit*father_SHS			0.205*	-0.072
			(0.121)	(0.087)
Father post-sec educ	0.006	0.003	0.073*	0.039
	(0.029)	(0.022)	(0.043)	(0.038)
Attrit*father_post-sec			0.206	-0.107
			(0.125)	(0.103)
Mother primary educ	-0.018	-0.008	0.042	-0.014
	(0.030)	(0.023)	(0.047)	(0.036)
Attrit*mother_primary			0.051	0.124
			(0.146)	(0.098)
Mother JHS educ	-0.008	0.005	0.063*	0.002
	(0.021)	(0.017)	(0.036)	(0.027)
Attrit*mother_JHS			-0.247**	0.078
			(0.107)	(0.062)
Mother SHS educ	-0.025	0.009	0.035	0.028
	(0.027)	(0.022)	(0.038)	(0.034)
Attrit*mother_SHS			-0.190	0.187*
			(0.135)	(0.104)
Mother post-sed educ	-0.018	0.015	0.118**	0.080**
	(0.032)	(0.027)	(0.055)	(0.040)
Attrit*mother_post-sec			-0.370**	0.150
			(0.149)	(0.122)
Number of siblings	-0.003	-0.001	-0.003	0.000
	(0.004)	(0.003)	(0.005)	(0.004)
Attrit*sibling			-0.024*	0.006
			(0.014)	(0.015)
Born in rural area	0.004	0.010	-0.014	-0.012
	(0.015)	(0.013)	(0.023)	(0.021)
Attrit*rural_born			0.043	0.021
			(0.066)	(0.050)
Constant	0.090	0.074	0.889***	1.166***
	(0.122)	(0.098)	(0.177)	(0.131)
Region and cohort FE	Yes	Yes	Yes	Yes
Observations	1737	2439	1737	2439
R-squared	0.027	0.023	0.215	0.213
Wald test			F(10,134) = 1.62	F(8,135) = 1.58
			p-value = 0.1081	p-value = 0.1367

Note : *** p<0.01, ** p<0.05, * p<0.1. Clustered at SHS level standard errors are in parenthesis.
Anne Duplantier (2021)

TABLEAU C.4 Tests of Wave Attrition - Substitution sample - Probit

	FGM test		BGLW test (Baseline value of)	
Dependent var.	Wave-attrition		Apply to uni	Attend at uni
	(1)	(2)	(3)	(4)
Applied to university	-0.039***			
	(0.015)			
Attended university		-0.050***		
		(0.014)		
Wave-attritors			0.259	-0.536*
			(0.602)	(0.290)
Actual WASSCE score	0.001	0.000	-0.016***	-0.022***
	(0.001)	(0.001)	(0.002)	(0.002)
Attrit*WASSCE_score			-0.013	-0.010
			(0.008)	(0.007)
Exp. uncondi returns	-0.000		-0.000	
	(0.000)		(0.000)	
Attrit*returns			0.000*	
			(0.000)	
Exp. max-min	0.004		0.011	
	(0.011)		(0.015)	
Attrit*max_min			-0.055	
			(0.058)	
Age	-0.001	0.002	-0.019***	-0.017***
	(0.004)	(0.004)	(0.007)	(0.006)
Attrit*age			0.013	0.026**
			(0.016)	(0.011)
Male	-0.003	-0.003	0.131***	0.098***
	(0.015)	(0.011)	(0.026)	(0.020)
Attrit*male			-0.152*	-0.100
			(0.083)	(0.066)
Father primary educ	0.042	0.048*	-0.049	-0.085
	(0.034)	(0.026)	(0.065)	(0.057)
Attrit*father_primary			0.000	-0.017
			(.)	(0.185)
Father JHS educ	-0.007	-0.009	-0.041	-0.060*
	(0.025)	(0.020)	(0.043)	(0.035)
Attrit*father_JHS			0.427**	-0.083
			(0.186)	(0.147)
Father SHS educ	0.031	0.019	-0.027	-0.020
	(0.028)	(0.023)	(0.041)	(0.038)
Attrit*father_SHS			0.410**	-0.135
			(0.192)	(0.142)
Father post-sec educ	0.007	0.003	0.063	0.019
	(0.028)	(0.023)	(0.042)	(0.037)
Attrit*father_post-sec			0.463**	-0.159
			(0.197)	(0.160)
Mother primary educ	-0.017	-0.008	0.045	-0.011
	(0.028)	(0.024)	(0.045)	(0.035)
Attrit*mother_primary			-0.032	0.254*
			(0.188)	(0.149)
Mother JHS educ	-0.009	0.005	0.063*	-0.007
	(0.019)	(0.017)	(0.036)	(0.027)
Attrit*mother_JHS			-0.398***	0.198*
			(0.129)	(0.120)
Mother SHS educ	-0.024	0.009	0.040	0.018
	(0.026)	(0.021)	(0.037)	(0.033)
Attrit*mother_SHS			-0.340**	0.272*
			(0.168)	(0.146)
Mother post-sed educ	-0.023	0.015	0.109**	0.056
	(0.033)	(0.027)	(0.051)	(0.036)
Attrit*mother_post-sec			-0.507***	0.299**
			(0.174)	(0.149)
Number of siblings	-0.003	-0.002	-0.003	0.000
	(0.004)	(0.003)	(0.005)	(0.004)
Attrit*num_siblings			-0.034**	0.012
			(0.017)	(0.014)
Born in rural area	0.003	0.010	-0.012	-0.007
	(0.014)	(0.012)	(0.022)	(0.020)
Attrit*rural_born			0.080	0.028
			(0.080)	(0.060)
Region and cohort FE	Yes	Yes	Yes	Yes
Observations	1737	2439	1728	2439
Pseudo R-squared	0.046	0.041	0.188	0.199
Wald test			Chi2(10) = 24.57	Chi2(8) = 11.61
			p-value = 0.0062	p-value = 0.1694

Note : *** p<0.01, ** p<0.05, * p<0.1. Clustered at SHS level standard errors are in parenthesis. Marginal effects are displayed. Anne Duplantier (2021)